Clinical Judgement in the era of Evidence Based Medicine

Flores Sepulveda, Luis Jose

Awarding institution:
King's College London

Clinical judgment in the era of Evidence Based Medicine

Submitted to the Department of Philosophy of King's College London,

in partial fulfilment of the requirements for the degree of

Doctor of philosophy

Luis Flores MD. MA.

Department of Philosophy, King's College London

September 2016

Abstract

"Evidence Based Medicine" (EBM) urges that medical recommendations be based on the best research evidence, rather than on clinical judgement. While I strongly endorse attention to relevant research evidence, I argue that the related downplaying of clinical judgement is a step backwards. This is because actual models of EBM encourage physicians to focus exclusively on research probabilities and so to neglect relevant information about patients. I call this feature of EBM the "Problem of Extra Information" (PEI), and contend that it leads to predictions and prescriptions based on the wrong probabilities.

The PEI has been largely neglected by EBM, which has construed the challenge of clinical care as a matter of developing better research evidence, and of reminding physicians to attend to patients' preferences and values. And although meritorious attempts have been made to connect research with individuals through sophisticated methodological improvements, these only address the PEI partially, and do not eliminate the need for clinical discretion.

In this dissertation I contend that, in response to the PEI, clinical medicine requires a more Discretionary Approach (DA). This approach recognizes that the objective probabilities that matter for clinical recommendations are those in the reference class defined by everything the physician knows about the patient, and argues that the central role for judgment in clinical practice is to estimate these probabilities.

So understood, the DA has two main advantages over the EBM approach: prudential adequacy and evidential flexibility. My defence of the DA consists of addressing criticisms of the role ascribed to judgment and clinical experience within this approach.

The final two chapters of this doctoral dissertation complement my arguments with two meta-analytical empirical studies: one which compares "therapeutic guidelines based on evidence" with "usual care" with respect to patients' outcomes, and another which examines the relative predictive performance of statistical models and physicians' judgment in the context of diagnosis and prognosis. These studies refute previous evidence cited against judgment and vindicate the plausibility of the Discretionary Approach to clinical care.

"Having turned a cold shoulder to the hoary (overused) notion of the intuitive "art of medicine," it might be the case that clinical judgment can now far more productively be seen as that critical faculty that is brought to bear when faced with uncertainty."

Goodman, 2003

Acknowledgments

I owe my gratitude to David Papineau for accepting me as a graduate student and providing me with insightful guidance over four years. I was very fortunate to have him as supervisor. His keen intellect and extensive philosophical experience made every meeting instructive and stimulating. I am also grateful for his intellectual flexibility and approach to scholarship, which allow me to pursue an interdisciplinary doctoral project that is fundamentally philosophical, but also involved the conduction of empirical research to inform clinical care.

This dissertation would not have been possible without the continuous love, closeness, support, affection, and criticism of my wife, Carolina. She is not only an intelligent and competent epidemiologist who shares many of my academic interests, but also (and fortunately) someone who usually looks at methodological and philosophical problems from a different perspective than mine. I am deeply grateful for her feedback and countless comments on every chapter of this thesis. Carolina also coauthored the article versions of the empirical work presented in chapters five and six.

I dedicate this dissertation to my father, whose constant support has been essential to complete my doctorate. He stimulated my interest in pursuing this project to address the questions that first emerged throughout the course of my medical training and further developed during my first years of practice as clinical psychiatrist.

I am also thankful for the assistance of my mother and my parents-in-law. Their support was fundamental in allowing my wife and me to carry out our respective doctoral projects, while simultaneously raising two wonderful children: Magdalena who is now five-years-old, and one-year-old Manuel José. My children have been a constant source of inspiration and meaning, even during the most difficult periods of this project.

Special mention deserves my friend and co-author Jonathan Fuller. He not only collaborated with the empirical research presented in chapters four and five, but also provided me with feedback on various topics of this thesis.

My earliest versions of the content presented in chapters 1, 3 and 4 were published in the article *"Therapeutic inferences for individuals"* (Flores, 2015) and presented in the conferences *"Interdisciplinary workshop in the philosophy of medicine: medical knowledge, medical duties*" (KCl, 2014)[1], *"The concept of clinical Judgement in the era of Evidence Based Medicine."(*VU, Amsterdam, 2015*) and "Diagnostic reasoning in Psychiatry: The case for analytical methods." (*AAPP, New York, 2014*)*

[1] Summary available in Bullock and Kingma, 2014

Finally, part of my thoughts on the problem of external validity and extra information were previously expressed, *"The Risk GP Model: the standard model of prediction in medicine"* (Fuller and Flores 2015), and *"Translating Trial Results in Clinical Practice: the Risk GP Model"* (Fuller and Flores, 2016).

Table of Contents

Introduction

Clinical medicine is medicine for individuals. It is neither an academic exercise devoted to the refinement of general medical knowledge nor a research activity focused on increasing our understanding of diseases as abstract entities. In the context of clinical practice medical evidence acquires significance insofar it is relevant and properly applied to particular patients.

This doctoral dissertation is about clinical medicine so understood. In particular, I am interested in how physicians should make diagnoses, prognoses, and therapeutic recommendations for individuals. I approach this topic from a theoretical and empirical perspective. I am interested in how physicians should apply evidence from various sources during clinical decision-making and the role of judgment during this process. I endeavour to make claims about the right probabilities for clinical recommendations, and the place of clinical judgment in the care of particular patients.

Scholarship concerning the logic underlying medical prescriptions, diagnostic expertise, and clinical judgment extends from the time of ancient medicine to relatively recent work on the application of decision theory to medical practice. As will be discussed in this dissertation, the official perspective on the question of how clinical medicine should be practised is currently dominated by the postulates of Evidence Based Medicine (EBM). One of the core ideas proposed by this movement, and one that has been fully embraced by the medical establishment, is that physicians should incorporate the best research results into practice, either directly or indirectly via evidence-based guidelines. Moreover, along with stressing the importance of research results, the EBM movement has questioned the roles traditionally attributed to clinical judgment and evidential sources such as clinical experience. Nowadays, therefore, the prevailing view is that optimal clinical recommendations are, almost without exception, recommendations warranted by the best research available.

The EBM movement undoubtedly has its merits. After all, who would deny that attention to valid research findings is important and potentially beneficial? Or who would like to be treated by physicians prescribing treatments supported by out-dated assumptions, unchecked reasoning, or even merely wishful thinking? And yet, the maxims of EBM can be taken to the extreme. And a number of supporters of EBM have taken them precisely there.

In his book The *Philosophy of Evidence-Based Medicine*, Jeremy Howick is not satisfied with endorsing the now uncontroversial view that clinical judgment alone is normally a poor basis for assessing general population-level claims about therapeutic efficacy. According to Howick, clinical judgment is too fallible a tool to be entrusted with *any* kind of evidential function during clinical care. Thus, he suggests that is not only in the context of the

generation of general medical knowledge that physicians should abandon clinical discretion. Physicians should also put their judgment to one side when comes to predictions and prescriptions for individuals.

Somebody might think that Howick's stance is idiosyncratic, and that few scholars would embrace such high levels of scepticism over the capabilities of clinical judgment. But this is hardly the case. The idea that *biased clinical judgement* is in part responsible for the so-called "*unwarranted variation in healthcare*" is accepted by prominent supporters of EBM and it is part of the rationale supporting a model of standardised clinical care where physicians are obliged to comply with evidence-based guidelines (EBGs). The line of thought is as follows: If two patients have the same condition and share similar preferences, how could it be right for them to receive different treatments? In line with this, David Eddy –a prominent figure in the camp of EBM– argues that, if one of these two patients receives a treatment supported by Randomised Controlled Trials and the other does not, someone must be guilty of an error of judgment and someone is receiving suboptimal care. Thus, in the eyes of those who are keen to promote standards of practice based on EBGs, the issue of non-compliance with validly obtained research results does not seem to deserve extended debate: preferences and values may be taken into consideration, but departures from EBGs for evidential reasons are unacceptable.

This commitment to a strong link between EBGs, standards of practice, and quality measures is hardly surprising, since it fits naturally with the hierarchical arrangement of evidential sources launched by the EBM movement. If recommendations from EBGs are supposed to be distillations of the best research evidence, then, departure from such recommendations is likely to be supported by low quality evidential sources (*personal clinical experience* or *clinical judgment* itself). Given this, it is unsurprising that physicians are urged not to exercise their judgment and depart from recommendations from EBGs—indeed if they choose to do so they risk accusations of suboptimal care or worst malpractice lawsuits.

Furthermore, it is not difficult to see why EBM policy makers are against clinical judgment and why they would like to keep it as restricted as possible. In the view of most of them, increasing the room for clinical discretion would be like opening a Pandora's box–permitting recommendations based on anecdote and worsening clinical care. So, from the perspective of the strongest believers in EBM, improvements in care will come from one direction: from research to practice. Patients need more and better research, more EBM practioners, and more physicians complying with EBGs. The less room for clinical discretion the better, because the right clinical care is, quite simply, care consistent with evidence-based interventions. In the most extreme, EBM implies the elimination of inferences modelled by clinical judgment and

informed by experience, in favour of management protocols exclusively based on high-quality research data. The presence of estimates from valid research override clinical judgment because what physicians have seen and learned about patients is likely to be biased, and their reasoning, even when it uses the best research as starting point, is likely to find too many exceptions to the applicability of clinical trials to individuals that are not justifiable according to EBM standards.

Of course, not all proponents of EBM think that medicine should dispense with clinical judgment altogether, or not at least for the present time. More balanced views within the EBM camp can be found, as there are those who do not think that the obliteration of the current gap between estimates for populations and estimates for individuals by ordinance is the best option available. In fact, the standard definition of EBM, which was formulated by David Sackett, one of the founders of this movement, seems to describe a version of EBM practice that is radically different to the extreme model of standardised care where the best research dictates what constitutes optimal practice. According to Sackett's definition, EBM practioners do have room for the exercise of clinical judgment; what is more, they should integrate the best research evidence with their clinical expertise, and do so in a conscientious way.

Sackett's description of EBM will be discussed in some detail in Chapter Two. At this stage let me simply observe that, once one pays close attention to actual models of EBM practice, it becomes apparent that, despite the mention of clinical expertise in definitions like Sackett's, EBM practitioners are encouraged and even obliged to practice in accordance to the best research available. The misalignment between how EBM is represented and how EBM is taught and practiced suggests that those who defend EBM against criticism by appealing to its standard definition do not want to recognize that it has become "Rubbish EBM" or, as JPA Ioannidis says, it "has been hijacked".

As I shall argue, the central problem with actual models of EBM practice is that they neglect what I call the Problem of Extra Information (PEI). This problem arises because the physician standardly knows more about the patient than that the patient belongs to some population for which research data are available, yet the EBM approach tacitly forces the physician to assume that this extra information can be neglected.

This problem emerges because, from my perspective, EBM has not formulated the challenge of clinical inference correctly. As we have seen in this Introduction, the practice of EBM centred on the identification and application of valid research. Thus, the main task of the EBM practitioner is either to search for and apply valid research or comply with recommendations based on valid research present in EBGs. This implies that clinical

judgment, if permitted to enter the scene, does so after valid research (or derivatively valid recommendations) have been clearly pinned down. This may seem trivial, but it is not. By focusing the physician's attention on the probabilities for which there is valid research available, the EBM approach turns the physician's attention away from the reference class defined by the physician's knowledge of the patient, and towards reference classes there are valid research data for. Thus, the EBM approach effectively dismisses the PEI because by the time clinical judgment enters the scene it is too late: the physician has already been directed to recommendations based on membership of a general diagnostic reference class for which reliable statistics are available, and so has been tacitly obliged to assume that the extra features present in her patient do not make any probabilistic different to the outcome of interest. One way to summarize this is to say that the EBM approach promotes recommendations based on the wrong probabilities, and this very fact means that, even if the standard definition of EBM mentions clinical judgment, this approach has not yet fully understood that one of the main roles of the physician is to address the PEI.

If what the physician is trying to do during the clinical encounter is to maximise benefit and minimise harm in her patient, then, the probabilities she ought to be interested in are the probabilities for patients exactly like the physician knows her patient to be. In terms of reference classes, my contention is that the right approach to clinical care should encourage the physician to look for the probabilities for the reference class conformed by everything she knows about the patient. These are, from a prudential perspective, the right probabilities to base clinical recommendations on and therefore, the probabilities the physician needs to make a judgment about. Since recommendations based on probabilities obtained via valid research can be but need not be relative to everything the physician knows about the patient, then, the probabilities supported by the EBM approach, even in its cutting edge version, need not be the right probabilities and it is reasonable to think that often are the wrong probabilities.

Defending this view is the task of this doctoral dissertation. I shall argue that the EBM approach as it is practiced and implemented is wrong in encouraging physicians to focus exclusively on research probabilities. In response to the limitations of the EBM approach, and in particular to its failure properly to address the PEI, I shall propose a more Discretionary Approach (DA). According to this approach, it is rational for the physician to base her clinical recommendations on the probabilities relative to the reference class defined by everything she knows about the patient. These are the right probabilities for clinical recommendations, and the role of clinical judgment is to address the PEI by estimating such probabilities.

My argument proceeds in five steps. The first step (Chapter 1) reconstructs the rationale of EBM recommendations in terms of a partially formal argument, which I think captures the

essential elements of this approach and permits me to illustrate how it works in practice. Using a clinical case, I shall illustrate the PEI and show how EBM recommendations lead to its neglect. The chapter explains the rationale behind focusing on what I call the "right probabilities", and shows how one of the advantages of the DA over the EBM approach is increased evidential flexibility.

The second step (Chapter 2) is devoted to defending my reconstruction of the EBM approach against a charge of straw man. I accept that there is a sense in which this charge appears to be justified by the standard definitions of EBM. However, by carefully examining two different models of EBM practice I provide evidence that suggests that the actual practice of EBM is more similar to my reconstruction than to EBM's official definition.

The third step (Chapter 3) considers the prospects of alternative ways of dealing with the PEI, which avoid my Discretionary Approach and its appeal to clinical judgment. I examine several research refinements –including Pragmatic RCTs, subgroup-analyses, and "n of 1 RCTs"– and I admit these are helpful in various ways. (Some of these are related to the PEI, while others help with standard difficulties related to statistical estimation and causal inference.) Then, I appraise "personalised" and "precision" medicine, which I accept can narrow the gap between research and individuals. Nonetheless, despite their benefits, I conclude that the research improvements in question ameliorate but do not solve the PEI, and therefore cannot be used to support the elimination of clinical discretion.

The fourth step (Chapter 4) further elaborates on the DA and examines potential worries about its application. First, the DA is distinguished from similar positions such as *Medical Particularism* and *Person Centred-Medicine*. Then, I defend it from various charges ranging from *lack of normative novelty* to *impracticality*. Finally, this chapter discusses the rationale behind the concern that the implementation of the DA might worsen clinical care, and argues that since supporters of EBM have overstated the negative consequences of clinical experience and physicians' limited cognitive abilities, and partially neglected the limitations of research evidence, well-designed and updated comparisons between the DA and the EBM approach are needed.

Finally, the fifth step (Chapters 5 and 6) presents the results of two systematic reviews, which shed light on the importance of implementing the DA. Chapter 5 provides the most comprehensive and updated summary on the relative performance of successfully implemented EBGs and "care as usual". The findings presented support the plausibility of the DA approach in the context of therapy. Chapter 6 then consist of a systematic review comparing the predictive accuracy of statistical models with that of physicians' judgment in the context of diagnosis and prognosis. This study also challenges previous findings in the

subject used by supporters of EBM to discredit the role of clinical judgment, and thus complements the findings presented in the previous chapter, vindicating the DA this time in the context of diagnosis and prognosis.

Chapter 1: EBM: a critical appraisal and a positive proposal

1.1. Abstract

The main aims of this chapter are to clarify the kind of recommendations endorsed by the Evidence Based Medicine (EBM) approach to clinical care, and to bring attention to one of its most important limitations. Although I strongly concur with the idea that attention to research evidence is crucial to sound clinical decision-making, I shall argue that the EBM approach is wrong in encouraging compliance with predictions and prescriptions that neglect part of what the physician knows about her patient. I shall call this feature of the EBM approach the Problem of Extra Information (PEI), and my main contention shall be that, to the extent that EBM disregards the implications of this problem for particular patients, it supports a type of clinical care based on the wrong probabilities.

In response to this problem, I shall argue that, if the physician's goal is to maximize expected utility for her patient, then her clinical recommendations need to emerge from a more discretionary approach. This approach hinges on the idea that the probabilities the physician ought to be interested in are the objective probabilities in the reference class determined by everything the physician knows about the patient, and that whenever the PEI arises, clinical judgment ought to be exercised so that recommendations are based on estimates of these probabilities.

1.2. The scope of clinical medicine

Clinical medicine is medicine for individuals. At its core there is the clinical encounter, which could be, roughly, described like this: There is a patient, maybe currently affected by some ailment, who visits a physician looking for information about her present situation and guidance as to what to do with it. When seeing the patient the physician gathers information about her and her current situation, and then makes various recommendations.[2]

If clinical practice is described as a set of predictions and prescriptions issued by physicians during the clinical encounter, then, the question arises: What kind of inferences should the rational physician make in this context? How should physicians arrive at rational clinical recommendations?

Nowadays, it is widely accepted that clinical medicine should be practised along the lines of Evidence-Based Medicine (EBM) [3] This movement has presented physicians with an intuitively compelling normative standard: clinical recommendations should ideally be based on the best research results. Why so? Because, as one of its founders puts it, research evidence of the highest quality is *"so much more likely to inform us and so much less likely to mislead us"* (Sackett et al. 1996. p.71). Thus, if supporters of EBM are right in that the best possible decision-making is decision-making warranted by high quality research, then it seems to follow that the rational physician ought to arrive at clinical recommendations following the rules of EBM.[4]

In this chapter I want to cast doubt on some aspects of EBM so understood. True, paying attention to research evidence –and in particular to relevant and carefully conducted research studies– is crucially important to improve the quality of medical care. But for all its importance, I shall argue that care based on the best research is not necessarily the best care for the individual. And, in fact, I shall argue that to think of the adequacy of clinical medicine in almost exclusive reference to the best research evidence is misplaced. Clinical medicine requires a more *discretionary approach*, an approach that recognises explicitly that the right probabilities for clinical inferences are the objective probabilities in the reference class defined by everything the physician knows about the patient; an approach that acknowledges

[2] Throughout this thesis my use of the term "the patient" should be understood broadly, as encompassing characteristics inherent to the patient and any aspect or feature of the clinical encounter.

[3] This claim is based on the fact that most medical schools (e.g. (Harvard Medical School, no date), (Cambridge Medical School, no date), professional associations –e.g. (American college of Physicians, no date), (Royal College of Physicians, no date) and the majority of health institutions, private and public (e.g. (NICE, no date) in United Kingdom and (NIH, 2016) in United States endorse the principles of EBM. For general sources on the impact of EBM see (Meats et al., 2009) and (Ioannidis, 2016).

[4] As the philosopher Michael Loughlin (2009) has complained, there are supporters of EBM who seem to have taken the reasonableness of this approach for granted and even have reacted with outrage or sarcasm to anyone who question the movement (e.g. Goldacre 2006; Isaacs and Fitzgerald, 1999).

without hesitation that valid research need not provide the physician with estimates of these probabilities, and an approach that promotes the exercise of careful judgment to determine the best recommendations for each individual.

The plan for this chapter is as follows. In section 1.3 I shall offer a brief discussion on basic aspects of expected utility theory, probability, and causation, so to clarify from the outset my position in the field of rational clinical decision-making. Section 1.4 is focused on the EBM approach. This section starts with a general reconstruction of this movement in order to make explicit the rationale behind evidence-based recommendations, and continues with a clinical case, which will help me to illustrate my main objection against the EBM approach: its lack of attention to the Problem of Extra Information. Then, this section ends by considering two initial replies from supporters of EBM (§ 1.4.3 and § 1.4.4), which will be useful in explaining why I think the EBM approach directs physicians' attention to the wrong probabilities. Finally, section 1.5 offers a more detailed account of the discretionary approach and two reasons for accepting that the probabilities in the reference class defined by everything the physician knows about the patient are the right probabilities for clinical inference: first, prudential adequacy and, second, evidential flexibility.

1.3. Rational clinical recommendations

1.3.1. Expected utility theory

Think again of the question I posed in the previous section: What kind of recommendations should the rational physician make in the context of clinical medicine? My strategy to analyse the answer put forward by EBM involves looking at the clinical encounter from the perspective of *expected utility theory (EUT)*.[5]

Consider a therapeutic prescription. When deciding what to prescribe the physician thinks of several possible outcomes (cure, quality of life, etc.), each of which has been assigned certain utilities. Then there are various therapeutic interventions available, and each of them has a certain probability to cause the outcomes of interest (which in turn depends on certain states of the world). Now, from the point of view of EUT, what the rational physician is going to do is to choose the treatment that *maximises expected utility*.[6]

[5] By expected utility theory I mean the traditional doctrine initiated by British utilitarian philosophers Jeremy Bentham and John Stuart Mill and further developed by John Von Neumann and Oskar Morgenstern. This theory has several formulations but, in essence, that the ideal rational decision-maker ought to pursue actions that maximize expected utility (Kyburg and Thalos, 2003).

[6] Given a set of therapeutic prescriptions open to the physician, $T_1, \ldots T_k$; a set of outcomes of interest $O_1, \ldots, O_m$, each with utilities $U(O_1), \ldots, U(O_m)$; and a set of conditional probabilities for O_i given therapeutic prescription T_j,

Obviously this is an idealization. But it will not matter for my purposes. EUT will be my starting point. It will provide us with a general normative framework, which I will use, first, to examine the EBM approach, and second, to lay down the foundations of the discretionary approach, the alternative approach I shall defend. As it will become apparent during the next sections, the main reason why EUT is important for my purposes is that it plays a central role in the prudential argument with which I will defend the role of judgment during clinical inference.

Nonetheless, being a practicing physician myself, I take very seriously arguments that complain that EUT is at best an idealisation of actual decision-making and so of limited relevance to actual clinical practice (Gilovich et al. 2003). In response to such concerns, the last two chapters of this dissertation (Chapters 5 and 6) present the results of a reanalysis of systematic reviews and a meta-analytical summary of evidence that shed light on the potential consequences of giving physicians more room for the exercise of their clinical judgment, and thereby provide valuable insights about physicians' capacity to get near to the normative ideal envisaged by EUT.

Moreover, it is also important to clarify from the outset that my main concern shall be with *probabilities* not *utilities.* The importance of utility functions in clinical medicine is difficult to overestimate, but there are two reasons why I shall not focus on them in this dissertation. First, although accurate estimates of both utilities and probabilities are required to maximize expected utility, the challenges presented by each of these functions are largely independent: even if utilities were perfectly known, difficulties with probabilities may result in inappropriate recommendations and vice-versa. The second reason why I shall concentrate on probabilities rather than utilities is that utility assessments have already received some attention in the medical literature. This is particularly so in the context of therapy, where the model of shared-decision making has brought attention to the importance of considering patients' preferences and values (e.g. Siminoff, 2013; McCartney et al. 2016). By contrast, the probabilistic rationale of physicians' clinical inferences, and specifically, the question of what kind of probabilities predictions and prescriptions should be based on, has been largely ignored by medical scholars. [7]

$P(O_1|T_1)$,…, $P(O_m|T_k)$; then, the therapeutic prescription that maximizes expected utility is the T_j for which is $\sum_{i=1\ldots m} U(O_i)\ P(R_i|T_j)$ is maximum (Jeffrey, 1983).

[7] My focusing on probabilities will be reflected in the analysis of cases where utilities are fixed so that the goal of maximizing expected utility be attained by recommending the action with the highest probability to cause the outcome of interest. Note, however, that this is only to facilitate exposition, for I do not intend to convey the idea that the rational physician should always recommend the action with the greatest probability to cause the outcome

1.3.2. Clinical recommendations: predictions and prescriptions

The components underlying clinical recommendations require elaboration. The practice of clinical medicine can be seen as an inferential process leading to series of predictions and prescriptions pertaining to individual patients (Card and Good, 1971). In this thesis the term clinical recommendations will encompass both *predictions* and *prescriptions* for particular patients, and since the focus of my analysis is restricted to clinical medicine, my use of these terms should not be extended to populations unless otherwise specified.

Note also that throughout the course of the three fundamental medical tasks: diagnosis, prognosis and therapy, physicians' predictions and prescriptions intertwine in complex ways (Murphy, 1997). For this reason, it will be helpful to specify how these terms will be understood. My usage of the term *prediction* will be restricted to *diagnostic* and *prognostic* tasks, and therefore circumscribed to inferences and subsequent recommendations which are aimed at detecting concurrent but yet unknown outcomes or forecasting future outcomes without causing them.[8] The term *prescription*, on the other hand, shall be reserved for recommendations involving active interventions aimed at *causing* the outcome of interest in the context of therapy.[9]

1.3.3. Causal and non-causal correlations

Since therapeutic recommendations typically involve interventions aimed at modifying the pathological process causing the patient' state, I will take it that *rational prescriptions* should be normally based on relevant causal correlations. By demanding attention to causal correlations rather than simple correlations, I am making explicit my commitment to *causal decision theory* in the context of medical therapy. In doing so, I am assuming that when physicians prescribe medications they should not be merely interested in *increasing the probability of the outcome of interest* but rather in *causing the outcome of interest in a probabilistic sense*.[10] On the other hand, since *diagnostic* and *prognostic recommendations*

---

of interest. Such recommendation only holds under the assumption that utilities make no difference, which is not true in many real clinical circumstances.

[8] For example, in diagnostic contexts, a prediction can be about the presence of a certain underlying disease based on a set of symptoms (in this case, a diagnostic prediction is identical to a diagnostic hypothesis). And, in prognostic contexts, a prediction can be about the future occurrence of an event of interest (e.g. acute myocardial infarction) without therapy.

[9] Of course, in common usage, the term "prescription" has sometimes been used in the context of diagnosis (e.g. when physicians "prescribe" a diagnostic test) or also to describe actions not aimed at causing the outcome of interest in the context of therapy (e.g. when physicians "prescribe" that the patient simply wait until the symptoms fade out without taking any particular therapy). Nonetheless, it will simplify my exposition to distinguish between (i) *diagnostic and prognostic predictions*, and (ii) therapeutic prescriptions.

[10] Note that I am aware that in many situations the treatment that increases the probability of the outcome will also be the treatment that causes the outcome in a probabilistic sense, but since this need not be the case it is worth

are concerned with prediction rather that causation, I shall take it that *rational predictions* need not be (though they can be) based on causal correlations.

The practical importance of the distinction between causal and non-causal inferences in the context of different clinical tasks can be illustrated with an example. Consider the relationship between an opportunist infection such as oesophageal candidiasis (OC) and the diagnosis of Human immunodeficiency virus (HIV). Many cases of HIV infection are first detected because the patient seeks medical attention due to symptoms related to OC. Since opportunist infections arise almost exclusively in the presence of immune deficiency, OC is considered a strong predictor of HIV infection (Patton et al. 1999). So, although OC does not cause HIV infection, a diagnostic prediction of HIV and subsequent diagnostic actions (carrying out specific HIV tests) in virtue of the presence OC are perfectly reasonable. By contrast, although the identification of OC is typically followed by the diagnosis of HIV infection, the fact that the former does not cause the latter makes it unreasonable to infer that a patient with HIV infection will recover if her OC is successfully treated with antifungal agents. So, whereas relevant non-causal correlations often suffice for reliable diagnostic and prognostic predictions, simple correlations might lead physicians astray when comes to therapeutic prescriptions. [11]

Note, in passing, that, even if the merits of causal decision-theory over evidential decision theory involve heated discussion among philosophers (Peterson, 2009; Papineau, 2006), it does not seem to be a contentious issue in the context of medical practice. For one thing, medical researchers and clinicians alike are now familiar with the notion of *confounding* and accept that this kind of factors has to be controlled for as much as possible in the context of therapy.[12] For another, most medical practioners accept as uncontroversial the distinction between (a) *risk and prognostic factors*, which are interpreted as *simple correlations* for

---

stating upfront that my normative account of therapeutic decision-making demands attention to causal probabilities whenever possible.

[11] Note the distinction between (a) the causal effect of the act of predicting and (b) a prediction based on a factor that causes the outcome of interest. When I say that diagnostic and prognostic tasks are not normally *aimed* at causing the outcome of inference I had in mind (a). But when I say that diagnostic and prognostic predictions can but need not be based on causal correlations I have in mind (b). Of course, with respect to (a), there are interesting cases where the act of forecasting unintentionally causes the outcome of interest via some psychological or psychosomatic mechanism. These kinds of mechanisms are not uncommon in certain patients (e.g. prediction of relapse in drug addicts, or prediction of symptoms' recurrence in patients with psychosomatic disorders). However, these are special cases, which do change the fact that most diagnostic and prognostic tasks involve actions that are causally independent of the outcome of interest (e.g. the act of requesting a chest radiography does not cause a patient's pneumonia, this action only changes the probability of detecting it). Moreover, regarding (b) let me stress that while there are situations where physicians predict outcomes using causal factors – for example, the prediction of cervical cancer by identifying the presence of Human Papilloma virus– most diagnostic and prognostic predictions are based on simple correlations.

[12] Evidence of an increased interest in methods to deal with confounders not only includes current attention to randomized controlled trials, but also the development of various techniques to deal with confounders in the context of non-randomized studies (Vineis, 1997; Listl et al., 2016; Hernán and Robins, 2006; Nichols, 2007).

purely predictive purposes, and (b) *causal factors and mechanisms*, which are included as part of the rationale behind therapeutic interventions (Knottnerus, 1995; Riley et al. 2013; Andersen, 2012). So, it is agreed across the board that therapeutic prescriptions need to be based on causal correlations, while robust diagnostic procedures or reliable prognostic forecasts need not be backed up by causal inferences.[13]

1.3.4. Subjective and objective probabilities

Finally, before proceeding with my reconstruction of the EBM approach, it will be useful to be explicit about the kind of probabilities that will occupy our attention. According to traditional decision theory (e.g. Jeffrey, 1983) a *subjectively rational physician* ought to maximise expected utility from the point of view of her *personal probabilities* or *subjective degrees of belief*. But even if the physician is successful at doing this, it is not difficult to imagine situations where *subjectively rational prescriptions* would not be objectively advisable. Suppose a physician needs to decide between two available treatments for the patient, who is currently affected by a gout attack. However, this physician is misinformed and in consequence estimates that treatment A (say, colchicine and ibuprofen) will cause recovery with probability 0.3 and the alternative treatment B (say, rest without medication) will cause the same outcome with probability 0.7. If we assume that the patient's preferences are neutral between the treatments, then this physician will *rationally* recommend her patient treatment B. And yet, should we think that this recommendation is advisable when the patient's objective probability of recovery is, say, 0.8 with treatment A and 0.4 with treatment B? Obviously, whenever *personal probabilities* do not match the relevant *objective probabilities* there is little point in being subjectively rational (Papineau, 2006). So, I shall take it as uncontentious that patients need physicians to choose prescriptions that will maximize their probability of recovery in an objective sense. Or, to put it even more explicitly, that the rational physician ought to attend to the right *objective* probabilities.

In this respect, I suspect that the intuition that physicians' practice should be relevantly connected to *objective probabilities* is one of the main thrusts behind the Evidence Based

[13] Having said that, two remarks are in order. Firstly, the fact that simple correlations can lead physicians astray in the context of therapy does not imply that the only information useful during therapeutic decision-making is information whose causal relevance to the outcome of interest has been established via valid research designs (e.g. Randomized Controlled trials). There are many genuine causal correlations supported by experience, whose effect size is so large that it would be both epistemologically unnecessary and unethical to carry out RCTs to rule out potential unknown confounding (in the classic example, the use of parachutes to avoid death related to gravitational challenge (Smith and Pell, 2003). Furthermore, without taking things to the extreme, it is reasonable to suppose that physicians can sometimes use their own experience to detect potentially relevant factors, which if have enough biological plausibility could be attributed a causal role and taken into account during clinical decision-making. Secondly, my commitment to causal decision theory does not imply that the output of standard research methods developed for purposes of causal inference (again, RCTs) is sufficient to ensure the adequacy of therapeutic prescriptions for individuals. As I shall explain during the next sections, in the context of therapy attention to general causal correlations from research should be complemented with appropriate consideration of potential interaction effects present in the patient in question.

Movement (EBM). And I must stress that this is an intuition with which I am happy to agree. However, I do not think the EBM movement is directing physicians' attention to the right objective probabilities. This is a matter of practical importance, for although I accept that paying attention to the valid research evidence is important for decision-making, I do not think that the right objective probabilities for clinical recommendations are always the probabilities for which valid research estimates are available.

In the next section I will offer a general reconstruction of the EBM approach, which along with a clinical case will illustrate how EBM encourages recommendations based on what, I shall argue, are the wrong objective probabilities.

1.4. Evidence-based clinical recommendations

1.4.1. The EBM argument

Let us consider an admittedly ideal situation. Suppose there is a population defined by the presence of a certain diagnosis D. Assume the physician has access to updated research evidence about the efficacy of certain medications in this population. This evidence is of the highest quality according to EBM rules.[14] It consists of a set of randomized controlled trials, which were conducted to compare the efficacy of medication A to that of medication B – the two treatments currently available for patients with diagnosis D. Suppose that the findings indicate that, on average, medication A has a higher probability to cause the outcome of interest (e.g. recovery R) than medication B. And, suppose further that the physician has good reason to regard these findings as both internally and externally valid: samples were representative of the underlying population D, randomization was not violated, outcomes were well-chosen, measurements were unbiased, treatment protocols and follow-up periods were realistic, and so on.

Imagine now that the physician has a particular patient p, and is sure the patient has ailment D. That is, let us pretend that the relevant diagnostic tests are perfect in terms of sensitivity and specificity [15], and so there is certainty as to the fact that patient p has the ailment in question. Finally, to complete this idealized scenario, assume that the physician has confidently established that the outcome of interest for patient p is "recovery" and that the patient is indifferent to other utilities related to the therapies available (e.g. economical costs, adverse effects, etc.).

[14] By EBM rules I mean hierarchies (or levels) of evidence published by institutions such as the Oxford Centre for EBM (Phillips et al., 1998; updated by Howick, 2011) or described in standard EBM textbooks (e.g. Guyatt et al., 2008; Strauss et al., 2011). Rules of evidence are described in more detail in Chapter 2 (§ 2.4.1.3). For a thoughtful and detailed examination of the hierarchical rules of evidence the reader is referred to Blunt (2015).

[15] Readers not familiar with these terms are advised to consult Akobeng (2007).

This hypothetical situation provides us with an excellent opportunity to introduce and examine the rationale of the EBM approach to clinical inference. For, in a situation just like the one described (or close to it), this approach tells us that the physician ought to recommend medication A.

In order to see what exactly is going on here it would be helpful to be explicit about the logical steps underlying this recommendation. Of course, such endeavour is not without problems: just as different arguments can lead to the same conclusion, the EBM rationale may be represented in different ways.[16] Nonetheless, a useful way to think about this is in terms of the following argument:

i) $p \in D$

ii) $P_D(R|A) > P_D(R|B)$

∴) A is the right therapeutic prescription for p

where "p" denotes a particular patient, "D" a given target population, "R" stands for recovery, and A and B denote mutually exclusive medications.[17]

Several remarks can be made about this argument, but let me start with what I think is the simplest but most interesting among them: although persuasive and perhaps popular among many practitioners of EBM, this argument is invalid. By this, I simply mean that even if the physician were perfectly certain of the truth of the premises, the conclusion would not follow.

This may be familiar ground to many philosophers, but since it is not for most physicians allow me to unpack this argument in more detail. The problem with this argument does not lie on the presence of false or implausible premises, but rather it has to do with the relationship between the premises and the conclusion. A closer look at the content of the premises will clarify this.

---

[16] Note that I am aware that not every author who considers himself a supporter of EBM will be committed to the same kind of prescriptions. Nonetheless, although I acknowledge divergence of though within EBM, I think that the reconstruction of the logical basis of the recommendations *normally attributed* to the EBM movement it is possible. This is so because, as I will argue in chapter 2, both EBM teachings and implementations strategies share a commitment to EBM rules of evidence, which underlies physicians' intuitive understanding of what counts as an evidence-based recommendation.

[17] Note that, as I explained previously in this chapter (§ 2.1), in order to focus on the logic of probabilistic inference this argument assumes that the patient's outcome of interest R is fixed and that the patient is indifferent to further utilities. It is only given these assumptions that the EBM argument can offer a conclusion about the right recommendation rather than merely about probabilities. So, the reader may well take the conclusion as equivalent to: *"If p is exclusively interested in R then A is the right therapeutic prescription".* Note also that in the context of evidence of ideal RCTs, the difference in the conditional probabilities showed in the second premise should be given a general causal interpretation such as "*in population D, on average, medication A has a higher probability to cause recovery than medication B*" (See, Papineau, 1994, 1985, 1989; also Cartwright, 2007).

The first premise represents what we might call a 'diagnostic hypothesis′. It simply asserts that the patient belongs to a set of subjects defined by the presence of a given medical condition, in this case ailment D. This is a perfectly plausible premise, which standardly arises during the exercise of clinical medicine. Of course, in real contexts physicians do not rely on foolproof diagnoses, but this merely indicates that a premise of this kind may be more or less certain depending on the patient in question, among other considerations, but there is nothing in it that makes it necessarily false. Thus, this diagnostic hypothesis seems to be a reasonable premise for supporters of EBM to rely on.

Let us now turn our attention to the second premise. As with the previous premise, this second premise does not seem questionable by itself. Here what it is being asserted is simply that a specific intervention has a higher probability to cause the outcome of interest than a relevant alternative on average in a certain population. Of course, in practice physicians rarely face an ideal situation where the evidence available is so strong that any concerns about the appropriateness of statistical inference (estimation), causal inference (confounding), and external validity (extrapolation) do not arise, but even if in real situations the aforementioned issues can be problematic, there is no principled reason to think that general efficacy claims are not practically possible.[18]

So, if the premises seem plausible, why is it that I claim that the conclusion does not follow? Is it not exactly in situations of this kind: where the patient's diagnosis is correct, the evidence seems both strong and relevant to the patient, and preferences and values have been properly accounted for, when it is reasonable to conclude that the application of the best research evidence determines the right prescription? Why is it that I claim that this EBM recommendation does not follow? An example inspired in a real situation suggested by Brian Hurwitz (Hurwitz, 2013) will be useful to illustrate the problem I have in mind.

1.4.2. The case of Mr. Smith and the problem of extra information

To continue with our useful idealization let us consider first a simplified clinical scenario A: Dr Jones is a physician who considers himself as an EBM practitioner. She wants to establish the best treatment for her patient, Mr Smith, who is currently affected by 'acute conjunctivitis'. Assume that the clinical picture is clear so the diagnosis is well justified. Now, Dr Jones, being the dutiful physician she is, knows very well that she should search for updated evidence of the highest quality to establish whether "a short course of antibiotics" or

[18] Let me clarify immediately that the main reason why at this stage I am setting aside problems underlying general efficacy claims is not because I think these are unimportant (Fuller and Flores, 2015, 2016; and Flores 2015), but rather because I want to bring attention to a further problem which remains even if statistical estimation, confounding, and extrapolation to target populations were not a problem. We will return to these issues in chapter 3 (§ 3.4), when discussing current methods to refine EBM recommendations.

"observation without antibiotics" is better for Mr Smith. Given that Dr Jones does not have much time to perform a systematic review and then critically appraise a long list of potentially relevant studies, she might opt to comply with the recommendations of evidence-based guidelines[19]. In this respect, current guidelines indicate that "acute conjunctivitis" is a self-limiting condition, in which topical antibiotics are not effective (Jefferis et al. 2011, Sheikh and Hurwitz, 2006; Steeples and Mercieca, 2012).

Now, given a solid guideline recommendation, backed up by 'gold standard' methods, and assuming that the preferences of Mr Smith are not an obstacle, an EBM practitioner like Dr Jones would not hesitate to conclude that Mr Smith should not receive antibiotics, and that this properly evidence-based advice does not deserve further revision. And yet, regrettably, since Dr Jones' individualized prescription is backed up by an inadequate rationale, it may be wrong in many circumstances, which will become apparent if we consider a slightly more realistic clinical scenario B.

Assume now that during the clinical interview Dr Jones records some extra information about Mr Smith. For example, she learns that Mr Smith's daughter has been recently affected by a serious case of acute conjunctivitis in which a particular bacterium was isolated. Or, alternatively, suppose that when asking for personal antecedents Dr Jones traces back a history of repeated sexually transmitted diseases (RSTD), or perhaps she learns that Mr Smith has recently returned from a trip to a suburban, poor area in Africa.

I take it that Dr Jones should not ignore this information. This is not because the information will necessarily change Mr Smith's diagnosis, for after considering this information there is still a sense in which the diagnosis remains acute conjunctivitis. Rather, it is because this information may make a difference to the clinically relevant objective probability of recovery for Mr Smith (Steeples and Mercieca, 2012; Postema et al.1996; Garland et al. 1995).

But then, how could Dr Jones take into account this information and simultaneously follow EBM rules? Does not Dr Jones knows that the best research evidence available indicates that antibiotics do not work for patients with acute conjunctivitis, and that Mr Smith has acute conjunctivitis, and so the right thing to do is to withhold antibiotics? If the EBM approach is correct, and the right prescriptions are prescriptions supported by the best evidence available, then it seems that Dr Jones should not waste her time paying attention to additional information. After all, the premises required by the EBM argument have already been met. Dr Jones knows that Mr Smith has acute conjunctivitis, and according to EBM guidelines, the

[19] The two alternatives mentioned here refer to the four steps model of EBM practice ("4S model", see chapter 2 (§ 2.4.1) and the EBM guidelines model of practice ("EBGs model", see chapter 2 (§ 2.4.2).

right thing to do in this case is not to give antibiotics. Thus, it seems that the EBM approach does not leave much room for Dr Jones to revise the probabilities that should guide her treatment of Mr Smith in the light of what she has learned about him, although it seems clear that, in truth, ignoring such information may result in the wrong prescription.

With this more realistic clinical scenario in mind, it becomes apparent that neither absolute certainty that Mr Smith belongs to the diagnostic category 'acute conjunctivitis', nor the availability of reliable research conducted on patients with this diagnosis are enough to ensure that the EBM prescription is the right one for Mr Smith.

Let me emphasise, in passing, that the problem with this EBM recommendation has nothing to do with a potential conflict of utilities (or "preferences and values" in the EBM jargon). The question of what utilities Mr Smith attaches to the available prescriptions is of crucial importance, but certainly different from the question of how likely is that each of the prescriptions available will cause the outcome of interest for Mr Smith.

So, to summarize: what is the trouble with the EBM approach? Put simply, the problem is that this approach does not account for the fact that the physician might know more about the patient than that the patient belongs to a population for which there are valid research data, and that irrespective of the presence of further information available to the physician at the time of decision-making, the EBM approach standardly encourages (and sometimes obliges) the physician to act on the probabilities for which valid research estimates have been obtained. This is problematic because, in real clinical encounters, physicians are likely to identify features that might make a difference to the probabilities that should guide treatment but for which valid clinical trial results are unlikely to be found. In the case of Mr Smith, EBM recommends Dr Jones to base her inferences on the best research data, but such data only exists for patients with acute conjunctivitis *tout court*, not for a patient like Dr Jones knows Mr Smith to be, that is, for patients with acute conjunctivitis and the characteristics Dr Jones learned about him during the clinical encounter (e.g. RSTD, recent trip, and so on).

This problem will be denoted the Problem of Extra Information (PEI). It is worth emphasizing that it arises whenever the physician has knowledge that locates the patient in a more fine-grained reference class, by which I mean an extensionally narrower but informationally richer reference class.

So described, the PEI has been largely disregarded by supporters of EBM. As my reconstruction of the EBM argument illustrates, the emphasis of this approach has been on recommendations based on validly obtained general efficacy claims (second premise), plus

reminding the physician that the individualisation of research findings requires attention to the patient's preferences and values.

In the rest of this chapter I shall argue that the PEI calls for a revision of the EBM approach. For insofar as physicians are advised to base their decisions on research estimates that ignore available information, EBM encourages them to try to identify the *wrong* objective probabilities to guide treatment. In response to this problem, I shall formulate an alternative approach. This approach holds that as the right probabilities for clinical inference are the objective probabilities in the reference class based on everything the physician knows about her patient. Hence, it emphasises that probability estimates from valid research may but need not be estimates of the right probabilities, and that EBM should not encourage physicians to assume that any extra-information available about the patient is probabilistically irrelevant. This approach will be denoted the discretionary approach (DA). Although I certainly think that rational physicians ought to attend to valid research evidence, this should not prevent them aiming to estimate the objective probabilities that should actually guide their treatment of each of their patients.

Before proceeding, however, it will be worth to address two immediate replies to the challenge raised by the case of Mr Smith.

1.4.3. Is this a straw-man?

Supporters of EBM may rely on several strategies to address the challenge raised by the case of Mr Smith. An obvious alternative would be to deny that the argument sketched in section 3.1 truly represents the EBM approach. The EBM movement, this rebuttal goes, does in fact leave space for clinical judgment, and therefore it gives physicians sufficient manoeuvrability to take into account additional information so to avoid misapplications of research results. Physicians who apply the best research evidence in the way Dr Jones did are not exemplars of EBM practitioners, since a truly EBM practitioner would have considered all available data to change her prescription.

EBM's advocates have the right to raise this objection because, up to now, I have not offered specific evidence to back up my reconstruction of EBM. An adequate reply to this objection requires a detailed account of how exactly EBM obliges physicians to apply valid research evidence to particular patients, and what the exact place of judgment within the EBM's system is. However, given that the answers to these questions require contextual information

and an extended argument, I shall address concerns about a potential misrepresentation of the EBM movement in the next chapter.[20]

For the moment, I can anticipate that although my reconstruction of EBM removes some nuances so to avoid unnecessary complications, it still captures the kind of practice that follows from standard models of EBM practice. While it is true that the most common definition of EBM (Sackett et al. 1996) seems to permit the exercise of clinical judgment, I shall maintain that the very presence of prescriptions supported by high-quality research data generates a normative pressure that in most cases ends collapsing almost completely the space for judgment and thereby makes EBM practice vulnerable to the problem of extra information (PEI).

More importantly, even if there is a sense in which the methods used to implement EBM recommendations permit the exercise of clinical judgment, I would still argue that, in practice, the space assigned to clinical judgment in this context does not really help the EBM approach to avoid the PEI. For, insofar as supporters of EBM are truly committed to hierarchies of evidence, which is one of the core organizing principles of this movement, they will have to admit that inferences informed by low-ranked evidence are not welcomed, and, because of this, the role of judgment will be normally restricted to that of preventing blatant misapplications of valid research evidence under exceptional circumstances. And while I would be prepared to admit that this sort of gatekeeper function for clinical judgement is better than ignoring clinical judgment altogether, I would still argue that the EBM approach is misguided, for the application of judgment should not be restricted to a limited set of evidential sources but rather it should be kept flexible so that the physician to be able to estimate, as best as possible, the right probabilities for each patient.

1.4.4. The problem of outliers

A different EBM response to the case of Mr Smith would be to deny its significance. Supporters of EBM might say that a particular case showing that the EBM approach leads to the wrong recommendation has no argumentative force against its general reasonableness.

This idea can be expressed as follows. Given certain probability distribution of recovery for patients with acute conjunctivitis, nobody should be surprised by the existence of subjects for whom the probability of recovery with and without antibiotics is very different from the

[20] This is mainly because I need space to explain that there is a disconnection between the way in which EBM is presented to physicians in the official "integrative" definition provided by Sackett and colleagues (1996) and the methods by which EBM is actually taught and implemented: the "4S model" and the "EBM guidelines model" (§ 2.41 and 2.4.2).

average. However, this is no reason to conclude that the EBM treatment is not the right intervention for Mr Smith. For, even if the EBM prescription to withhold antibiotics will be wrong for some patients with acute conjunctivitis, it remains true that this prescription will be the right one for most of them.

Correct, it is a fairly common situation that research data indicates that the majority of patients with certain diagnosis will recover with treatment, but the physician does not know whether the particular patient in front of him belongs to this majority or belongs to the minority who will not recover (let us call this last group "outliers"). And while it is true that any particular member of the target population might turn out being in the group of outliers, I agree with supporters of EBM that this does not necessarily turn an inference based on average research data unreasonable.

Why is it that I do not see outliers as a problem for the reasonability of clinical inferences? As I pointed out previously (§ 1.3.1), from the perspective of expected utility theory, a reasonable prescription is not a prescription that "ensures" that the outcome of interest will occur, but rather a prescription that maximizes the probability of causing the outcome of interest. In consequence, I take it that if the patient in question turns out having an outcome that differs from average, one could simply acknowledge that both the patient and the physician were victims of "bad luck". However, by no means do I think that one should conclude, on this ground, that the physician's prescription was unreasonable. After all, the uncertainty inherent to the physician's prescription –that is, the inescapable possibility that the patient might end in the group of outliers– has been already addressed by the principle of rationality that tells the physician that she ought to maximize expected utility. So, from this perspective, the possibility of unlucky outcomes is no argument against the reasonability of the recommendation to prescribe the treatment that, on average, has a highest probability to cause the outcome of interest.

Now, does this imply that supporters of EBM are right when they say that the case of Mr Smith tells us nothing about the adequacy of the EBM approach? By no means, and let me emphasise why: the case of Mr Smith was not meant to illustrate that outliers are a problem for the EBM approach, but rather, as I said before, to bring attention to a distinct problem: the Problem of Extra Information (PEI).

The problem of outliers and the PEI may seem similar in the surface, but they constitute very different kinds of worries. The problem of outliers is, as I explained, a worry about the unavoidable possibility that probability-based prescriptions might not in the end bring about the desired outcomes. The PEI, on the other hand, is a concern about the amount of information the physician is taking into account at the time of decision-making. Thus, the

case of Mr Smith should not be read as providing evidence that the EBM approach is inadequate *because* the EBM recommendation to withhold antibiotics turned out to be wrong for Mr Smith, but rather as illustrating that the EBM approach is wrong because it supports a prescription that ignores part of the information available, and thereby tacitly assumes that the extra features Dr Jones knows about Mr Smith do not cause any difference to his probability of recovery. As my reconstruction of the EBM rationale illustrates (§ 1.4.1), the judgment responsible for taking into account this information is not part of EBM's inferential algorithm, and it is for that reason that this approach is vulnerable to the PEI.

So, to repeat, whether the case of Mr Smith ends up being an outlier is beside the point. The moral of the case of Mr Smith is not that in the particular situation illustrated things went (or could have gone) wrong, but rather that the EBM prescription is directing attention to probabilities in the wrong reference class. Concretely, the EBM prescription is incorrect because, from the point of view of Dr Jones, Mr Smith is not just a patient with acute conjunctivitis; Mr Smith is a patient with acute conjunctivitis and, say, a history of RSTD. Hence, my objection is that by endorsing prescriptions supported by valid research, which tacitly assume that the extra information available to the physician is probabilistically irrelevant, the EBM approach is guilty of directing physicians' attention to the wrong objective probabilities.[21]

1.5. The Discretionary Approach

So far, the central point I have made is that in its emphasis on ensuring prescriptions based on valid research estimates, the EBM approach forces physicians' attention towards the objective probabilities for which reliable statistics have been obtained. Let me, from now on, identify these probabilities using the term *EBM probabilities*. However, although I accept that attention to valid research is important, in my view encouraging recommendation based solely on EBM probabilities is misguided because such recommendations implicitly assume

[21] Notice that if the real clinical situation were such that the physician only knew that Mr Smith has Acute Conjunctivitis, then I would be prepared to endorse the EBM recommendation, provided that the EBM argument (§ 1.4.1) were modified as follows:

i) $p \in D$

ii) $P_D (R|A) > P_D (R|B)$

iii) This is everything the physician knows about p

∴) A is the right therapeutic prescription for p

This new argument still is not deductively valid, but I take it as inductively reasonable for it makes explicit the fact that the problem of extra information has not arisen. Nonetheless, as the more realistic version of the case of Mr Smith showed, it seems illusory to suppose that there will be many actual clinical situations where this additional premise could be satisfied. Normally, physicians know more about their patients than the mere fact they belong to a general diagnostic reference class like "acute conjunctivitis". Such kind of limited knowledge is commonly used as a pedagogical simplification but it is uncommon in real clinical contexts.

that any additional information about the patient is probabilistically irrelevant to the outcome of interest.

In this section I shall articulate in more detail an alternative to the EBM approach: the Discretionary Approach (DA) to clinical practice. This approach consists of two main elements. The first one is that the right probabilities for clinical recommendations are the probabilities in the reference class defined by everything the physician knows about the patient (*DA probabilities*); and the second element is that the physician ought to make a judgment about what these probabilities for each particular patient are.

Since concerns about the practical consequences of the Discretionary Approach will be addressed in subsequent chapters (Chapter 4, 5 and 6), at this point I shall concentrate on two reasons for accepting DA *probabilities* as the right probabilities of interest during clinical decision-making: namely, prudential adequacy and evidential flexibility.

1.5.1. Prudential adequacy

What are the objective probabilities of interest if the physician wants to maximize expected utility for her patient? My answer to this question is that the objective probabilities of interest are the probabilities in the reference class defined by everything the physician knows about the patient (DA probabilities). However, if this answer is accepted, it follows that the EBM approach is in trouble, for this approach encourages them to try to identify the *wrong* objective probabilities.

Let me put this idea in abstract terms so to make it explicit. Suppose some *EBM probability*, that is, a probability for which a valid research estimate is available, tells the physician that in some reference class R the probability of recovery for patients with ailment D given treatment T is k. Assume now that the physician knows that a particular patient with D is not just R but also X and Y. And assume further that, according to the physician's considered judgement based on his medical knowledge, the probability of recovery in the reference class defined by R, X and Y is not k but another number j. Then it is easily perceived that it would be irrational for the physician to assess the potential benefit of treatment T for this particular patient by using the probability k rather than j.

The case of Mr Smith clearly illustrates that there are situations where the EBM approach encourages physicians to act irrationally, in the sense that it supports recommendations based on probabilities that ignore part of what they about the patient. For if Dr Jones follows the EBM advice she is obliged to put her judgment to one side, and tacitly assume that factors such as a history of Repeated Sexually Transmitted Diseases (RSTD) are probabilistically

irrelevant to Mr Smith's probability of recovery. But, of course, it is plain that, from a prudential perspective, if Dr Jones wants to maximize expected utility for Mr Smith, the potential virtue of a treatment with antibiotics for Mr Smith ought not to be judged on the basis of the probability of recovery in the reference class defined by the presence of acute conjunctivitis but by the presence of acute conjunctivitis and a history of RSTD. It is this later probability the probability of interest for Dr Jones, and it is precisely because the EBM approach does not provide him with an estimate of this probability that she has to exercise her judgment to estimate, as best as possible, its value. [22]

1.5.2. Evidential flexibility

In this final section I would like to show how the DA allows—indeed encourages– physicians to make sensible use of a wide range of evidential sources. As I have said, and I shall demonstrate in the next chapter, the EBM approach ties the adequacy of physicians' recommendations to certain evidential sources via EBM rules of evidence. By contrast, the DA does not force physicians' attention towards any particular type of evidential source. The right recommendation, from the perspective of the DA, is right not because it is supported by certain kind of evidence (e.g. well-conducted RCTs), but because it is the recommendation based on estimates of the right probabilities, that is, the probabilities for the reference class defined by everything the physician knows about the patient. Of course, the DA endorses the idea that the physician would ideally like to obtain as good as possible an estimate of the right probabilities, but unlike the EBM approach the DA recognises explicitly that a perfectly valid evidential source may lose most or even all its appeal if it does not deliver estimates of the right probabilities. And, in the other end, the DA has no problem in recognizing that there are many situations where the right recommendations for particular patients are recommendations supported by what EBM rules consider low-quality evidential sources.

[22] Given the way I defined *DA probabilities* an immediate question arises: How much additional information should the physician gather about the patient before issuing a clinical recommendation? Should the physician invariably try to know more facts about the patient, with a consequent deferment in predictions and prescriptions? The proof provided independently by Frank P. Ramsey (1990) and Irvin J. Good (1967) is relevant to answering this question. These authors' result show that, whenever the cost of acquiring extra information about the patient is negligible (near zero), the expected utility of delaying the clinical recommendation until additional information is acquired is never less that that of issuing the clinical recommendation immediately. Of course, in real-life clinical situations gathering additional information about the patient does typically involve costs (including a variety of harms, time, money and so on), and so the rational physician will apply her judgment to weigh these costs against the extra expected utility of waiting on the further information before making the clinical recommendation. Notice, however, that since it is prudentially rational for the physician to guide all these calculations by DA probabilities, they are merely special cases of the superiority of the DA over the EBM approach. In "Probability as a Guide to Life" Helen Beebee and David Papineau (1997) develop a compelling argument to take acting on *relative probabilities* –equivalent to *DA probabilities*– as a primitive fact about prudential rationality, which underlies the Ramsey-Good result that agents ought to arrange to act on probabilities relative to more information, whenever this is not too costly.

In virtue of its evidential flexibility the DA approach leaves enough room for the exercise of clinical judgment. And, because of this, it can accommodate with ease common criticisms against the EBM approach, such as: (i) EBM's understatement of the problem of external validity of RCTs (Cartwright, 2007, 2011; Worrall, 2007), (ii) EBM's general lack of attention to knowledge about mechanisms (Clarke et al. 2014; Dragunilescu, 2016), as well as, (iii) EBM's effective dismissal of clinical experience (Greenhalgh and Worrall 1997; Greenhalgh, 2002; Greenhalgh et al. 2004, Elwyn et al. 2016; Tonelli, 2001; Upshur, 2005).

Since I shall consider some of the problems related to these criticisms against EBM in chapter 3 (§ 3.3.3) I shall now limit myself to explain very briefly how the DA is better prepared than the EBM approach to deal with the aforementioned objections.

1.5.2.1. External validity

One way to think about the problem of external validity is as a situation where it is unclear what actual target population research results are representative of, and for this reason neither physicians nor policy-makers know to what extent the findings, which may well be impeccable in terms of statistical estimation and control for confounding, are applicable to the intended target populations (e.g. diabetic patients in primary care).

Despite the fact that supporters of EBM have sometimes recognized that the samples used in clinical trials (and clinical trials' set ups in general) do not permit the extrapolation of findings to target populations (Guyatt el al. 2011), EBM's standard attitude towards the applicability of research designs ranked at the top of hierarchies of evidence (e.g. well conducted RCTs) is one in which the burden of proof is on the side of those who want to prove that research is not applicable.[23] This attitude, which has received several names, for example, *"simple extrapolation"* (Fuller, 2016) and *"extrapolation unless"* (Stegenga, 2015), emphasises that research evidence should be considered applicable *by default* and that only strong reasons could revert the extrapolation of research findings to target population in real clinical settings. These types of approach to applicability are particularly surprising because, while conventional RCTs randomize treatment allocations, in these trials samples are not normally randomly drawn from target populations.[24]

As I said earlier, the problem of external validity is different from the PEI. While the former is related to the task of establishing what is the target population clinical trials are

[23] See (§ 2.4.1.4) on "applying evidence" in the context of the 4S model of EBM practice.
[24] Notice that it is not uncommon that physicians infer an actual target population by relying on the arbitrary stipulations of researchers (in the case of primary research) or guideline-makers (in the case of EBM guidelines). In this regard, a review by Jonathan Fuller found that guideline-makers typically present guidelines as applicable to broad target populations despite the fact that primary studies are often conducted with highly selected samples (Fuller, 2013).

representative of, the latter has to do with using everything the physician knows to determine the right reference class for probabilities for each patient. However, one of the by-products of paying attention everything the physician knows about the patient (DA probabilities) rather than paying attention to the probabilities for which valid estimates have been obtained (EBM probabilities), is that it is less likely that physicians neglect the problem of external validity. This is because the DA, unlike the EBM approach, does not push physicians to restrict their attention to valid research evidence, and therefore eliminates the pressure to obtain some valid probability estimate to back up the right clinical recommendations and practice medicine in accordance with EBM rules.

Of course, supporters of EBM might complain that the DA neither ensures that the problem of external validity will not be neglected, nor makes its solution any less difficult. But, even so, it is undeniable that the DA may help physicians not to forget that sound clinical recommendations demand judgment in the light of all available information, and such judgment not only needs to be exercised to address the PEI, but also the problem of external validity. As Jonathan Fuller (2016) suggests when arguing against EBM's "*simple extrapolation"* approach to external validity, reasons in support of and against extrapolation should be taken into account, and it seems plausible to think encouraging attention to everything the physician knows about the patient may help him to find both types of reasons.

1.5.2.2. Mechanisms

The claim that basic scientific research on biological mechanisms and derivatively the so-called "physiopathological rationale" are not as reliable sources of medical knowledge as it was once thought is one of the hallmarks of the EBM movement (e.g. EBM working Group, 1992; Howick, 2011; Howick et al. 2013; Bluhm and Borgerson, 2011). In fact, one of the crucial elements of the campaign by which EBM came to occupy its current position as the standard approach to clinical care was its capacity to provide empirical data suggesting that, had physicians attended to randomized controlled clinical trials (RCTs), rather than to basic research about mechanisms, thousands of lives could have been saved.[25] Thus, current EBM rules emphasise that evidence of mechanisms has limited epistemological value in the context of therapy by downgrading its place in hierarchies of evidence, or excluding it completely (Bluhm, 2005).

Nonetheless, as several authors have pointed out (e.g. Feinstein and Horwitz 1997), supporters of EBM moved too quickly from the sensible claims that evidence about mechanisms can sometimes be misleading and that an exclusive reliance on

[25] This evidence was persuasively presented in the form of "horror stories". Specific cases can be found in Howick (2011, p.122-153) and Abel and Koch (1999).

physiopathological rationale can be problematic, to the idea that any knowledge about mechanisms and associated inferences lead to questionable recommendations, which are almost invariably inferior to those based on evidence from valid clinical trials. Furthermore, in their enthusiasm for the virtues of RCTs, supporters of EBM embraced the controversial idea that causal claims supported from such trials provide physicians with efficacy claims that need not be backed up by (or at least be consistent with) plausible mechanisms.[26]

Critics of the EBM movement argue, rightly in my view, that there are situations where research about mechanisms can provide physicians with evidence that not only complements but may even countervail the applicability of findings from RCTs to individuals (Clarke et al. 2013, Clarke et al. 2014; Russo, 2012). More precisely, attention to relevant mechanisms can help physicians in a number of situations including assessments of external validity as well as when it comes to determining whether findings applicable to populations are applicable to particular individuals (Clarke et al. 2013). So, unlike the EBM approach, the DA has ample room to incorporate evidence from mechanisms, for, from this point of view, this kind of evidence is not considered to be of inherently poor quality.

On the other hand, it is worth noting that the DA is also sensitive to the reasonable point that physiopathological reasoning can sometimes be misleading. This is the reason why this approach does not link the adequacy of physicians' inferences to specific evidential sources. According to the DA, physicians are responsible to answer the question of when and to what extent evidence of mechanisms should be considered on a case-by-case basis, for this question is essentially, a matter of clinical judgment. While for some patients the estimation of the right probabilities will be strongly influenced by relevant mechanistic evidence, such evidence need not have such effects in every patient. This is a matter of judgment, which, to repeat, is concerned with the estimation of the probabilities in the reference class defined by everything the physician knows about the patient.

At this point, supporters of EBM might complain that by appealing to physician's judgment the DA approach is merely avoiding an explicit confrontation with the problem of misleading mechanistic evidence. But the fact remains that where EBM implicitly forces physicians to ignore mechanistic evidence, the DA tells them that they must pay attention to it whenever the estimation of the right probabilities demands it.

1.5.2.3. Clinical experience

[26] Russo and Williamson have made a compelling case for attending to relevant mechanisms for purposes of causal inference in medicine (Russo and Williamson, 2007). See Weber (2009) and Claveau (2012) for interesting critical comments, and Illari (2011) for a lucid clarification of the Russo-Williamson thesis.

Another respect in which the DA fares better than the EBM approach has to do with the possibility to attend to the relevant clinical experience. Even though the way in which medical training has and continues to be structured suggests that medical knowledge developed through practical experience remains central to physicians' performance (Ludmerer, 2004; Huddle and Heudebert, 2008; Epstein et al. 2013), it is undeniable that the EBM movement has helped to cast doubt upon both physicians' capacity to acquire knowledge from regular clinical experience (Howick, 2011), and also upon the validity and reliability of such knowledge in the first place (Greenhalgh et al. 2014).

We will consider the way in which EBM is taught and implemented in chapter 2 (§ 2.4) but presently let me point out that, after the arrival of EBM the epistemological function of clinical experience fell into disrepute (Tonelli, 1998, 1999; Tanenbaum, 1993, 2012). From the perspective of supporters of EBM, the idea that physicians ought to attend to research evidence is as much justified by the virtues of research evidence as it is by the pitfalls of standard clinical experience as a source of general medical knowledge (Eddy 1990a-c, 2005; and Howick, 2011). Thus, hierarchies of evidence that exclude clinical experience –or include it implicitly and ranked at the bottom, embedded in the notion of expert judgment, are no more than a crystallization of EBM's standard attitude towards this kind of information.

True, I concur with supporters of EBM that assessments of efficacy based on uncontrolled clinical experience can be misleading, and also that physicians, as other humans, have several cognitive features that might lead them to mistaken inferences (Dawson and Arkes 1987; Elstein and Schartz 2002; Bornstein and Emler, 2001; Casarett, 2016). Furthermore, personal clinical experience, which is often characterised by a small number of patients, which might be inaccurately recorded and poorly recalled, can be particularly misleading for purposes of generating medical knowledge in the form of efficacy claims for general reference classes, or knowledge about general risk or prognostic factors.

However, having said that, I take it that only the most extreme among EBM's advocates would deny that personal experience can be useful to improve physicians' inferences in the context of clinical medicine (that is, when inferences are directed towards and predictions and prescriptions are restricted to particular individuals). In this context clinical experience can help physicians in different ways (Tonelli, 2006, 2011). It might help physicians to make sense of idiosyncratic features present in the kind of patients that visit their clinics. More precisely, experience might help physicians to determine whether certain features are or are not probabilistically relevant to the outcome of interest, or it might be the source of specific estimates, estimates for fine-grained classes of patients, for whom research data is unlikely to

be found (e.g. patients of specific age groups, who develop their conditions under local social circumstances, or who have uncommon comorbidities).

It is noteworthy, then, that just as the EBM approach had little room to allow for the evidential use of mechanistic reasoning, it also has little room for the incorporation of knowledge from clinical experience (Greenhalgh and Worrall, 1997). For, once more, if physicians want to avoid a major conflict with hierarchies of evidence, they will have to resist the temptation to use knowledge from experience to refine or replace probability estimates that come from more reliable evidential sources, such as RCTs. So, while within the EBM framework the use of clinical experience is normally interpreted as an action that decreases the quality of the resulting inferences or as a desperate resource in the absence of valid research evidence, the DA, by contrast, demands the exercise of clinical judgment in the light of information from clinical experience the patient's characteristics or situation so require it. According to the DA, therefore, there will be patients for whom clinical experience will be the type of evidential source underlying optimal inferences and patients for whom information from clinical experience will not be useful.

Finally, as with the limitations of mechanistic reasoning, given that the scope of the DA is restricted to particular patients, this approach is sufficiently qualified to acknowledge that clinical experience does not offer solid grounds to substantiate therapeutic or predictive claims for broad populations. So, the benefits of attending to clinical experience can be maximized in the benefit of well-selected individuals without encouraging physicians into unwarranted generalizations.

1.6. Conclusions

In this chapter, I reconstructed the rationale behind the EBM approach to clinical medicine, and argued that it promotes recommendations that are vulnerable to the PEI. This problem consists of encouraging recommendations based on the probabilities for which valid research estimates have been obtained, which implicitly forces the physician to assume that any extra information about the patient is irrelevant to the outcome of interest.

I argued that this makes the EBM approach inadequate. If the goal of the physician is to maximize expected utility for her patient, then the information she possesses about him at the time of decision-making (or the additional facts she could learn about the patient at a negligible cost before making a recommendation) should not be tacitly disregarded.

In response to this problem, I proposed that clinical medicine requires a more discretionary approach (DA). This approach acknowledges the importance of valid research findings, but it

is also sensitive to all information available to the physician, regardless of its source. This is because the DA is premised on the idea that the right probabilities for clinical recommendations are those in the reference class defined by everything the physician knows about the patient, and not just by the part of this information for which valid research estimates are available. Finally, the DA stresses that physicians have an active role in ensuring the soundness of clinical recommendations, for whenever the PEI arises the role of clinical judgment is to estimate, as best as possible, the right probabilities.

1.7. References

Abel U. and Koch, A. (1999). The role of randomization in clinical studies: myths and beliefs. *J Clin Epidemiol.* 52(6):487-97.

Akobeng, A.K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatr*. 96(3):338-41.

American College of Physicians. [no date]. *"Clinical Guidelines and Recommendations from the American College of Physicians"* Retrieved 27 December, 2015, from https://www.acponline.org/clinical-information/guidelines

Andersen, H. (2012). Mechanisms: what are they evidence for in evidence-based medicine? *J Eval Clin Pract.* 18(5):992-9.

Beebee, H. and Papineau, D. (1997). Probability as a guide to life. *Journal of Philosophy.* 94(5):217-43.

Bluhm, R. (2005). From hierarchy to network: a richer view of evidence for evidence-based medicine. *Perspect Biol Med.* 48(4):535-47.

Bluhm, R. and Borgerson, K. (2011). *Evidence-based medicine*. In *Philosophy of medicine: Vol. 16. Handbook of the philosophy of science.* D.M. Gabbay, P. Thagard, ans J. Woods (eds). Amsterdam, Elsevier. pp. 203-238.

Blunt, C.J. (2015). *Hierarchies of evidence in evidence-based medicine.* PhD Thesis, The London School of Economics and Political Science.

Bornstein, B.H. and Emler, A.C. (2001). Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *J Eval Clin Pract.* 7(2):97-107.

Cambridge Medical School. [no date]. "*Prospective medical students"* Retrieved 27 December, 2015, from https://www.medschl.cam.ac.uk/education/prospective-students/

Card, W.I. and Good, I.J. (1971). Logical foundations of medicine. *Br Med J.* 1(5751):718-20.

Cartwright, N. (2007). Are RCTs the Gold Standard? *BioSocieties.* 2(1):11-20.

Cartwright, N. (2011). A philosopher's view of the long road from RCTs to effectiveness. *Lancet.* 377(9775):1400-1.

Casarett, D. (2016). The Science of Choosing Wisely — Overcoming the Therapeutic Illusion. *N Engl J Med.* 374:1203-5.

Clarke, B., Gillies, D., Illari, P., Russo, F. and Williamson, J. (2013). The evidence that evidence-based medicine omits. *Prev Med.* 57(6):745-7.

Clarke, B., Gillies, D., Illari, P., Russo, F. and Williamson, J. (2014). Mechanisms and the Evidence Hierarchy. *Topoi.* 33:339–60.

Claveau, F. (2012). The Russo-Williamson Theses in the social sciences: causal inference drawing on two types of evidence. *Stud Hist Philos Biol Biomed Sci.* 43(4):806-13.

Dawson, N.V. and Arkes, H.R. (1987). Systematic errors in medical decision making: judgment limitations. *J Gen Intern Med.* 2(3):183-7.

Dragulinescu, S. (2016). Mechanisms and Difference-Making. *Acta Analytica*. 1-26.

EBM Working Group. (1992). Evidence-Based Medicine: a new approach to teaching the practice of medicine. *JAMA.* 268(17):2420-5.

Eddy, D.M. (1990a). The Challenge. *JAMA.* 263(2):287-90.

Eddy, D.M. (1990b). Clinical decision making: from theory to practice. Practice policies --what are they? *JAMA.* 263(6):877-8, 880.

Eddy, D.M. (1990c). Practice Policies: Where Do They Come From? *JAMA.* 263(9):1265, 1269, 1272.

Eddy, D.M. (2005). Evidence-based medicine: a unified approach. *Health Aff (Millwood).* 24(1):9-17.

Elstein, A.S. and Schartz, A. (2002). Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*. 324(7339):729-32.

Epstein, A.J., Srinivas, S.K., Nicholson, S., Herrin, J. and Asch, D.A. (2013). Association between physicians' experience after training and maternal obstetrical outcomes: cohort study. *BMJ.* 346:f1596.

Elwyn, G., Wieringa, S. and Greenhalgh, T. (2016). Clinical encounters in the post-guidelines era. *BMJ.* 353:i3200.

Feinstein, A.R. and Horwitz, R.I. (1997). Problems in the "evidence" of "evidence-based medicine". *Am J Med.* 103(6):529-35.

Flores, L. (2015). Therapeutic inferences for individual patients. *J Eval Clin Pract.* 21(3):440-7.

Fuller, J. (2013). Rhetoric and argumentation: how clinical practice guidelines think.. *J Eval Clin Pract.* 19(3):433-41.

Fuller, J. (2016). *The new medical model: chronic disease and evidence-based medicine.* PhD Thesis, University of Toronto.

Fuller, J. and Flores, L.J. (2015). The Risk GP Model: the standard model of prediction in medicine. *Stud Hist Philos Biol Biomed Sci.*54:49-61.

Fuller, J. and Flores, L.J. (2016). Translating Trial Results in Clinical Practice: the Risk GP Model. *J Cardiovasc Transl Res.* 9(3):167-8.

Garland, S.M., Malatt, A., Tabrizi, S., Grando, D., Lees, M.I., Andrew, J.H. and Taylor, H.R. (1995). Chlamydia trachomatis conjunctivitis. Prevalence and association with genital tract infection. *Med J Aust.* 162(7):363-6.

Gilovich, T., Griffin, D. And Kahneman, D. (2003). *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge, Cambridge University Press.

Goldacre, B. (2006). *Objectionable 'objectives'* [Online]. The Guardian. Available at https://www.theguardian.com/science/2006/aug/19/badscience.uknews [Accessed 05 December 2015].

Good, I.J. (1967). On the Principle of Total Evidence. *The British Journal for the Philosophy of Science*. 17(4):319-21.

Greenhalgh, T. and Worrall, J.G. (1997). From EBM to CSM: the evolution of context-sensitive medicine. *J Eval Clin Pract*. 3(2):105-8.

Greenhalgh, T. (2002). Intuition and evidence--uneasy bedfellows? *Br J Gen Pract.* 52(478):395-400.

Greenhalgh, T., Kostopoulou, O. and Harries, C. (2004). Making decisions about benefits and harms of medicines. *BMJ.* 329(7456):47-50.

Greenhalgh, T., Howick, J. and Maskrey, N.; Evidence Based Medicine Renaissance Group. (2014). Evidence based medicine: a movement in crisis? *BMJ.* 348:g3725.

Guyatt, G., Rennie, D., Meade, M.O. and Cook, D.J. (2008). *Users' Guides to the Medical Literature: A Manual for Evidence Based Clinical Practice*. 2nd Ed. New York, McGraw-Hill.

Guyatt, G.H., Oxman, A.D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., Alonso-Coello, P., Falck-Ytter, Y., Jaeschke, R., Vist, G., Akl, E.A., Post, P.N., Norris, S., Meerpohl, J., Shukla, V.K., Nasser, M. and Schünemann, H.J.; GRADE Working Group. (2011). GRADE guidelines: 8. Rating the quality of evidence--indirectness. *J Clin Epidemiol.* 64(12):1303-10.

Harvard Medical School. [no date]. "*Fundamentals of Medicine"* Retrieved 27 December 2015, from http://hms.harvard.edu/departments/medical-education/md-programs/new-pathway-np/fundamentals-medicine

Hernán, M.A. and Robins, J.M. (2006). Estimating causal effects from epidemiological data. *J Epidemiol Community Health.* 60(7): 578-86.

Howick, J. (2011). Exposing the Vanities—and a Qualified Defense—of Mechanistic Reasoning in Health Care Decision Making. *Philosophy of Science.* 78 (5):926-40.

Howitz, J. (2011). *The Philosophy of Evidence-based Medicine.* Oxford, Wiley-Blackwell.

Howick, J., Glasziou, P. and Aronson, J.K. (2013). Problems with using mechanisms to solve the problem of extrapolation. *Theor Med Bioeth.* 34(4):275-91.

Huddle, T.S. and Heudebert, G.R. (2008). Internal medicine training in the 21st century. *Acad Med.* 83(10):910-5.

Hurwitz, B. (2013). Confessional ruminations on an ocular cure testimonial. *Atrium*. 11(3):12-5.

Illari, P.M. (2011). Mechanistic evidence: disambiguating the Russo–Williamson thesis. *International studies in the philosophy of science.* 25(2):139-57.

Ioannidis, J.P. (2016). Evidence-based medicine has been hijacked: a report to David Sackett. *J Clin Epidemiol.* 73:82-6.

Isaacs, D. and Fitzgerald, D. (1999). Seven Alternatives to Evidence-based Medicine. *BMJ.* 319(7225):1618.

Jefferis, J., Perera, R., Everitt, H., van Weert, H., Rietveld, R., Glasziou, P. and Rose, P. (2011). Acute infective conjunctivitis in primary care: who needs antibiotics? An individual patient data meta-analysis. *Br J Gen Pract*. 61(590):e542-8.

Jeffrey, R. (1983). *The Logic of Decision. 2nd Ed.* Chicago, University of Chicago Press

Knottnerus, J.A. (1995). Diagnostic prediction rules: principles, requirements and pitfalls. *Prim Care.* 22(2):341-63.

Kyburg, H.E. and Thalos, M (2003). Probability is the Very Guide of Life: The Philosophical Uses of Chance. Chicago, Open Court Publishing.

Listl, S., Jürges, H. and Watt, R.G. (2016). Causal inference from observational data. *Community Dent Oral Epidemiol.* 44(5):409-15.

Ludmerer, K.M. (2004). The clinical experience in medical education: past, present, future. *Mo Med*. 101(5):487-90.

Loughlin, K.M. (2009). The basis of medical knowledge: judgement, objectivity and the history of ideas. *J Eval Clin Pract*. 15(6):935-40.

McCartney, M., Treadwell, J., Maskrey, N. and Lehman, R. (2016). Making evidence based medicine work for individual patients. *BMJ.* 353:i2452.

Meats, E., Heneghan, C., Crilly, M. and Glasziou, P. (2009). Evidence-based medicine teaching in UK medical schools. *Med Teach*. 31(4):332-7.

Murphy, E.A. (1997). *The Logic of Medicine.* 2nd Ed. Baltimore, The John Hopkins University Press.

Nichols, A. (2007). Causal inference with observational data. *The Stata Journal*. 7(4):507–41.

NICE [no date]. "*Evidence search*" Retrieved 27 December, 2015, from https://www.evidence.nhs.uk/

NIH (2016). "*Evidence-Based Practice and Health Technology Assessment"* Retrieved 27 December, 2015, from https://www.nlm.nih.gov/hsrinfo/evidence_based_practice.html

Papineau, D. (1985). Probabilities and Causes. *Journal of Philosophy*. 82:57-74.

Papineau, D. (1989). *Pure. Mixed, and Spurious Probabilities and their Significance for a Reductionist Theory of Causation.* In *Minnesota Studies in the Philosophy of Science XIII: Scientific Explanation.* P. Kitcher and W. Salmon (eds). Minneapolis, University of Minnesota Press. pp. 307-48.

Papineau, D. (1994). The virtues of randomization. *British Journal for the Philosophy of Science.* 45:437–50.

Papineau, D. (2006). *The Roots of Reason: Philosophical Essays on Rationality, Evolution, and Probability.* New York, Oxford University Press.

Patton, L.L., McKaig, R.G., Eron, J.J. Jr, Lawrence, H.P. and Strauss, R.P. (1999). Oral hairy leukoplakia and oral candidiasis as predictors of HIV viral load. *AIDS.* 13(15):2174-6.

Peterson, M. (2009). *Causal vs. evidential decision theory*. In *An Introduction to Decision Theory*. Cambridge, Cambridge University Press. pp. 187-199.

Phillips, C.B., Sackett, D., Badenoch, D., Straus, S., Haynes, B. and Dawes, M. (1998). (updated by Howick, 2009) *"CEBM Levels of Evidence"* Retrieved 01 January, 2016, from http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/

Postema, E.J., Remeijer, L. and van der Meijden, W.I. (1996). Epidemiology of genital chlamydial infections in patients with chlamydial conjunctivitis; a retrospective study. *Genitourin Med.* 72(3):203-5.

Ramsey, F.P. (1990). Weight or the Value of Knowledge. *The British Journal for the Philosophy of Science*. 41(1):1-4.

Riley, R.D., Hayden, J.A., Steyerberg, E.W., Moons, K.G., Abrams, K., Kyzas, P.A., Malats, N., Briggs, A., Schroter, S., Altman, D.G. and Hemingway, H.; PROGRESS Group. (2013). Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med.* 10(2):e1001380.

Royal College of Physicians. [no date]. "*Mission and strategy*" Retrieved 27 December, 2015, from https://www.rcplondon.ac.uk/mission-and-strategy-0

Russo, F. (2012). Philosophy of medicine: between clinical trials and mechanisms. *Metascience*. 21(2):387-90.

Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International studies in the philosophy of science*. 21(2):157-70.

Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B. and Richardson, W.S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ*. 312(7023):71-2.

Sheikh, A. and Hurwitz, B. (2006). Antibiotics versus placebo for acute bacterial conjunctivitis. *Cochrane Database Syst Rev*. (2):CD001211.

Siminoff, L.A. (2013). Incorporating patient and family preferences into evidence-based medicine. *BMC Medical Informatics and Decision Making.* 13(Suppl 3):S6.

Smith, G.C. and Pell, J.P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ*. 327(7429):1459-61.

Steeples, L., and Mercieca, K. (2012). Acute conjunctivitis in primary care: antibiotics and placebo associated with small increase in the proportion cured by 7 days compared with no treatment. *Evid Based Med.* 17(6):177-8.

Stegenga, J. (2015). Measuring effectiveness. *Stud Hist Philos Biol Biomed Sci*. 54:62-71.

Straus, S., Glasziou, P., Richardson, W. and Haynes, B. (2011). *Evidence-based Medicine: How to Practice and Teach It. 4th Ed*. Edinburgh: Churchill Livingstone.

Tanenbaum, S.J. (1993). What physicians know. *N Engl J Med*. 329(17):1268-7.

Tanenbaum, S.J. (2012). Improving the quality of medical care: the normativity of evidence-based performance standards. *Theor Med Bioeth.* 33(4):263-77.

Tonelli, M.R. (1998). The philosophical limits of evidence-based medicine. *Acad Med.* 73(12):1234-40.

Tonelli, M.R. (1999). In defense of expert opinion. *Acad Med.* 74(11):1187-92.

Tonelli, M.R. (2001). The limits of evidence-based medicine. Respir Care. 46(12):1435-40; discussion 1440-1.

Tonelli, M.R. (2006). Integrating evidence into clinical practice: an alternative to evidence-based approaches. *J Eval Clin Pract.* 12(3):248-56.

Tonelli, M.R. (2011). Integrating clinical research into clinical decision making. *Ann Ist Super Sanita.* 47(1):26-30.

Upshur, R. (2005) Looking for rules in a world of exceptions: Reflections on Evidence Based Practice. *Perspectives in Biology and Medicine. 48(4):477-489*

Vineis, P. (1997). Proof in observational medicine. *J Epidemiol Community Health*. 51(1):9-13.

Weber, E. (2009). How probabilistic causation can account for the use of mechanistic evidence. *International Studies in the Philosophy of Science*. 23(3):277-95.

Worrall, J. (2007). Evidence in medicine and evidence based medicine. *Philosophy Compass* 2/6. 981–1022, 10.1111/j.1747-9991.2007.00106.x

Chapter 2: Rigid versus judicious Evidence-based practice

2.1. Abstract

In the previous chapter I reconstructed Evidence-based medicine (EBM) as an approach to clinical care that is almost exclusively concerned with aligning recommendations with findings from valid research evidence. I argued that this is problematic because it directs physicians' attention to probabilities that ignore part of the data available, and thereby generate what I called the Problem of Extra Information (PEI).

But my reconstruction of EBM is controversial. Advocates of this movement might contend that I misrepresented EBM practice, for the reconstruction I offered presents EBM as if it were rigidly tied to research findings, when in truth, the practice of EBM has been standardly defined and described by its founders as an integrative and judicious endeavour, which is sensitive to all available information.

This chapter argues that this latter understanding of EBM is misleading. True, I shall accept that this approach is officially presented to the public using EBM's standard definition. Nevertheless, I shall argue that this definition is an idealised portrayal, which lacks descriptive accuracy due to a clear disconnection with the ways in which EBM is actually taught and implemented. By the end of this chapter, I expect to convince the reader that if enough attention is paid to the normative pressure imposed by the rules of EBM, as well as by the system of care that has been built around EBM teaching tools and guidelines, it becomes clear that EBM practice is much closer to my "rigid" reconstruction than to its standard "judicious" definition.

2.2. Rigid Evidence Based Medicine

In the first chapter of this thesis I reconstructed the EBM approach in terms of a two-premise argument (§ 1.5.1). The first premise is a diagnostic premise, which identifies a particular patient as a member of a standard target population. The second premise is a general efficacy premise, where valid research provides physicians with an estimate of the probabilities of some outcome of interest for members of a standard target population. Once these two premises are met –and the patient's utilities are not an obstacle– I claimed that the EBM approach advices physicians to prescribe the treatment supported by the general efficacy premise (§ 1.5.2).

My reconstruction of EBM practice emphasised that in the presence of evidence of the highest quality (e.g. well-conducted clinical trials), the physician is advised to ignore additional information available about her patient. I called this the Problem of Extra Information (PEI). When this problem arises, EBM recommendations based on the best available research become recommendations based on the wrong probabilities for the individual in question. In addition, my account of EBM practice stressed that EBM rules deemphasise reliance on evidential sources such as clinical experience or basic mechanisms (§ 1.5.2). That is why, in the context of therapy, I represented the EBM approach as committed to the idea that the EBM practitioner should comply with EBM guidelines based on evidence of the highest quality even when extra information seems to countervail standard recommendations (§ 1.5.2). And, similarly, in the context of diagnosis, I claimed that the EBM practioner is advised to act on estimates from statistical prediction rules, regardless of the presence the PEI, and therefore does not allow him to refine predictions and improve subsequent recommendations (§ 1.6.1). In both situations, I claimed that EBM recommendations are guided by the general EBM principle that clinical practice ought to conform to valid research findings whenever the patient belongs to the target population for which such findings are available (and, of course, personal preferences are no impediment).

But this portrayal of EBM practice is debatable. Advocates of the EBM movement might protest that I did not represent EBM recommendations accurately, and therefore that I am guilty of setting up a straw man to attack. Furthermore, supporters of EBM might say that the inadequacy of my description of EBM is very easy to demonstrate, for it happens to stand in sharp contrast with the standard account of EBM (Sackett et al, 1996), according to which there is much more to EBM practice than the rigid application of valid research evidence to individuals.

Now, since Sackett represents EBM practice as a judicious and flexible endeavour, it is hard to see how EBM recommendations could be as deprived of judgment as my reconstruction

suggests. So, if supporters of EBM are right, and it is true that EBM's standard definition describes how EBM works in practice, then, it appears that they are correct in saying that my "rigid" account of EBM (REBM) is no more than a straw man.[27]

Nonetheless, I shall attempt to convince the reader that this line of reasoning is misguided. For, I will argue that, in spite of its apparent influential character, Sackett's standard definition has little or nothing to do with the actual practice of EBM. Supporters of EBM who defend its judicious character and flexibility using Sackett's definition as if were descriptively accurate, I will contend, are victims of naivety or denial. For if they pay enough attention to the way in which EBM is actually taught and implemented, they will have to acknowledge that core normative principles of EBM such as the hierarchical rules of evidence, and the presence of significant extrinsic incentives, end almost entirely removing the space for judgment as well as obliterating the epistemological import of information not backed up by valid clinical trials.

This chapter is organised as follows. In § 2.3 I shall look closely at EBM's standard definition so as to identify what kind of practice is implied by it. Then, in § 2.4, I shall turn my attention to two methods by which EBM is currently being taught and implemented: the 4S model (§ 2.4.1) and the EBM guidelines model (§ 2.4.2). I shall argue that each of these methods is designed to indoctrinate EBM practitioners in the rules of EBM, and via different mechanisms, to promote recommendations vulnerable to the PEI and to discourage an effective use of clinical judgment to arrive at the "right probabilities" for clinical inference (§ 1.5.1). Thus, I shall conclude that the actual practice of EBM is much more consistent with my reconstruction of EBM that with Sackett's account.

2.3. EBM's standard definition

In the article *"Evidence Based Medicine: what it is and what it isn't"* (1996) [28] leading proponents of EBM define this approach as *"the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients"* and indicate that *"The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research"* (p.71-2).

[27] It is important to stress from the outset that Sackett's definition (1996) is an attempt to describe what EBM *is* (and it *is not*), not an attempt to describe how EBM *ought to be* given such and such aims.

[28] Notice that there are several definitions of Evidence-Based Medicine. Michael Loughlin, (2008 and 2009) surveys some of them. However, I shall focus on Sackett's definition (1996) for three reasons. The first one is that critics and supporters of EBM have treated Sackett's definition as the official definition of EBM (e.g. Bluhm and Borgerson, 2011; Greenhalgh et al, 2014; Djulbegovic et al., 2009). The second reason is that, as I said earlier (footnote 28) Sackett's definition is supposed to be a descriptive definition, which allegedly tell us about EBM practice as it actually is as opposed to as it ought to be. The third reason is that, with more than 4.300 citations, Sackett's definition (1996) is by far the most popular representation of EBM in the literature (Web of Knowledge, accessed December, 2015).

Now, since this definition is supposed to be descriptively accurate, two questions arise: First, what do EBM recommendations according to Sackett's depiction of EBM look like? Second, are such recommendations consistent with those that follow from my rigid reconstruction of EBM?

In answering the first of these question one faces an immediate difficulty: the official description of EBM seems to be rather vague, and so it makes it difficult to know how EBM recommendations look like.[29]

On the one hand Sackett's references to a *"conscientious"* and *"judicious"* practice might make the reader suppose that the EBM approach is something similar to the DA. For start, if the word *"conscientious"* is interpreted as suggesting a degree of *"carefulness"* and "thoroughness" in the use evidence then it might suggest EBM practitioners have enough room for the exercise of judgment. Similarly, while the meaning of the word *"judicious"* is not clear-cut, *"judicious use of evidence"* in the context of Sackett's description of EBM and the claim that EBM is not *"cookbook medicine"* seem to convey the idea of a well-considered inferential process, which again might led someone to think that EBM practice encourages physicians to exercise their clinical discretion. After all, if EBM practitioners acted *"judiciously"* when using evidence, it seems unlikely that they would make recommendations that assume that any information about particular patients is probabilistically irrelevant to the outcomes of interest.[30]

However, other phrases in the official description of EBM indicate that Sackett seems to favour something closer to my rigid EMB. In particular his references to "explicit evidence" [31] and his emphasis on the importance of "clinical research" make it hard to understand him as supporting DA. Moreover, while there are passages where Sackett seems to describe EBM practice as a flexible and integrative endeavour or when calling for the use of clinical

[29] Ironically, the vagueness of EBM's definition has not been considered in any way problematic by some supporters of EBM, whom seem to take the virtues of the approach to obvious to deserve careful thinking about it (e.g. Goldacre, 2006)

[30] Notice, however that there are authors in the EBM camp, who have explicitly claimed that patients would be better off if physicians put their clinical judgment to one side and thereby suggest that a "judicious use of evidence" would be to do exactly that (e.g. Howick, 2011; Eddy, 1990a).

[31] If EBM practitioners aim at an "explicit use of evidence" as Sackett describes, then there is a sense in which they are likely to conform to my rigid reconstruction of EBM. The reason for this is the following. EBM practitioners who do not want to ignore the PEI will presumably have a hard time trying to account for their use extra information in cases where their judgment tells them that such information changes the probability of the outcome of interest. This is because the relevance of extra information and their influence in recommendations would normally be supported by evidential sources such as tacit knowledge or perhaps their own clinical experience. Even more clearly, to the extent that clinical judgment itself is interpreted not only as a process but as an evidential source of second order (as it often happens with the testimony of experts in legal contexts), one would think EBM practitioners that are willing address the PEI would find difficult to use "explicitly" evidential sources that are inherently "implicit").

expertise [32], this is effectively cancelled out by the insistence that both clinical experience and judgment are subservient to the rules of evidence (Howick, 2011; Stegenga, 2014).[33]

Thus, on the one hand Sackett's characterisation suggests that EBM practice is flexible. But on the other hand, there are places where he suggests the contrary. In the end, as Michael Loughlin has pointed out (Loughlin, 2008, 2009) EBM's description is so imprecise that, it would be fruitless to try to pin down exactly what message is conveyed by Sackett's characterisation of EBM. The more important issue is to try to understand those who cite this definition and teach EBM to clinicians. In the rest of the chapter I shall argue that the actual models of EBM practice do not really encourage the EBM practitioner to use evidence judiciously, and do not really allow her to integrate different sources of evidence such as clinical experience, as Sackett's definition seems to imply.

2.4. EBM practice as it is taught and implemented

In this section we will consider two methods by which the EBM movement has been put into effect. I will argue that, if attention is paid to these methods, it becomes clear that EBM practioners are encouraged to make recommendations almost exclusively based on the best research evidence. Thus, the way in which EBM is taught and implemented leads to a kind of practice, which is vulnerable to the PEI, leaves little room for clinical judgment and, in the end, is much more consistent with my reconstruction of EBM than with Sackett's official definition.

The first method we will consider is a teaching method: the "4S" model (Sackett 1986, 1989, 1993; Straus et al. 2005). On the surface, this teaching approach does not appear to entirely rule out the exercise of clinical judgment; however, careful examination reveals that exercise of judgment is discouraged and in practice confined in extent.

This method achieves this effect by focusing physicians' attention on valid clinical trials from the outset and promoting the idea that the default attitude towards these trials should be that results are generally applicable to any member of the intended target population. Clinical judgment, if exercised at all, plays the limited role of preventing blatant misapplications of research evidence at the end of an inferential chain that seems to be almost determined to conclude applicability. Furthermore, while attention to extra information about the patient is

[32] When describing the practice of EBM Sackett claims that it implies "integrating individual clinical expertise with the best available external clinical evidence from systematic research" (Sackett et al 1996, p.71).

[33] As Gordon Guyatt and colleagues, in the name of the EBM working group explain (2000) explain: "The hierarchy implies a clear course of action for physicians addressing patients problems–they should look for the highest available evidence from the hierarchy." (p.1293). Notice also that in chapter 4 (§ 4.4.3) we will consider what Howick (2011) describes as the position of EBM on clinical judgment in some detail, and that according to Howick clinical judgment ought not to be attributed any evidential role, not even in the context of inferences for individuals.

not explicitly forbidden, adherence to the rules of EBM, which work as organising principles of the 4S model, deemphasises it by problematizing the use of evidential sources such as clinical experience or knowledge about mechanisms.

The second method through which EBM practice has been implemented is composed by EBM guidelines (EBG model) (Eddy, 1990b-f; Grol and Grimshaw, 2003; Grimshaw and Russell, 1993; Grimshaw et al. 2006; Qaseem el al. 2010; Woolf et al. 1999; Flodgren et al. 2016). Unlike the 4S model this second method does not rely on physicians' ability to search for and subsequently apply the best research evidence. The basic building block of this model is composed of readily applicable recommendations supported by evidence selected by guideline-makers. In theory, the EBG model need not result in recommendations that neglect the PEI, for, even if recommendations are exclusively based on research evidence, physicians might be able to incorporate individual patients' idiosyncrasies if they have enough room to exercise their judgment. However, reasons related to the way this system of practice is structured make it unlikely. First, clinical discretion and attention to other sources of evidence is discouraged by the normative power of strong recommendations, which are presented as actionable rules vouched by valid research evidence. Second, adherence to EBG is positively rewarded by quality-improvement schemes (e.g. pay-for performance initiatives), which pay physicians practices when is consistent with EBM recommendations. And third, the fact that several EBGs are equated to "best practices" in policy and legal contexts makes physicians think twice before deciding to apply their judgment to pay attention to information whose relevance might only be justified by low-ranked evidential sources according to EBM standards.

2.4.1. The four step model of EBM practice

The four-steps model (or "4S" for short) is one of the standard educational approaches to teach and promote EBM practice. This model was first introduced by Sackett and colleagues (Sackett et al, 1985) and it has been reproduced, with minor modifications, in several EBM textbooks and teaching articles (e.g Rosenberg and Donald, 1995; Akobeng, 2005; Dans et al. 2008; Prasad, 2013). In essence, the "4S" provides the physician with a recipe to base clinical decisions on valid research evidence. The process consists of the following steps: *"ask, acquire, assess, and apply"* (Prasad, 2013 p.2). In the first step the physician has to formulate a clinical question; the second step involves searching for evidence to answer this question; in

step three the physician assesses the validity of the evidence retrieved; and finally, in step four, the evidence obtained should be applied to the patient.[34]

Then, one would ask, what kind of EBM practice should be promoted by the 4S model for it to be consistent with the recommendations that follow from Sackett's definition?

In the next sub-sections (§ 2.4.1.1 and § 2.4.1.2), I will show that, contrary to what one would expect for a teaching model congruent with EBM's official definition, the 4S model has been operationalized so as to promote recommendations almost exclusively based on research evidence and thus to ignore part of what the physician knows about the patient. The first three steps of this model encourage selective attention to certain types of research evidence and step four (§ 2.4.1.4), while apparently leaving some room for physicians to consider additional information about individuals, ends up encouraging the physician to assume that such information is not worthy of attention.

2.4.1.1. Step 1: Asking for evidence

In the 4 steps model (4S) the task of asking for evidence for purposes of clinical decision-making is standardly presented as requiring the formulation of "*a focused clinical question"* (Guyatt et al. 2000. p.1294; Dans et al. 2008). Clinical questions need to be focused since proponents of this model expect them to be "answerable" by research evidence (Richardson et al. 1995). Thus, the very first step of the 4S model immediately reveals that although this teaching approach has been presented to physicians as general strategy to learn EBM practice, in truth it is a model that fosters exclusive attention to research findings from the outset. To see this, it is instructive to pay close attention to the ***PICO method***, that is, the technique physicians have to apply in order to construct proper "answerable questions".

The acronym PICO stands for "Patients", "Intervention", "Comparisons", and "Outcomes". This method is sometimes presented as a method to "*facilitate finding the best evidence*" (McKibbon et al. 2002 p.16). But, in truth, it is a method designed to help physicians to find evidence from selected clinical trials, as opposed to sources such as basic research, let alone relevant data from clinical experience (Richardson et al. 1995).

The PICO process encourages the physician to focus on clinical trials via a series of abstractions inherent to its application. These abstractions are, up to a point, understandable, for they are necessary to connect questions about particular patients with research findings

[34] More recently a fifth step has been added, which involves evaluating the results generated by the application of evidence (Haynes, 2006). However, since this last step is not relevant for the points I shall make in this chapter, we will continue our discussion based on the 4S model.

about general populations from the literature. But these abstractions come with a significant cost, namely the danger that the physician ignores potentially relevant information about the patient that was available from the outset.

To see this, consider the element "P", which normally refers to a type of "patient" or a "problem" in the PICO process. In regard to this element the treating physicians is advised to ask questions such as *"How would I describe a group of patients similar to mine?*" (The University of Illinois at Chicago, 2016). Of course, in answering questions of this kind, it is normal to expect some level of simplification. As I said, in order to establish a connexion between a single individual and research data, some simplifications have to be made. Obviously, one hopes that most of those simplifications will not be too detrimental to the inferential process. However, what is remarkable about the PICO process in the 4S model is that it encourages the physician to focus on a single purpose, namely, to extract away the information that makes the patient different from the research subject. This implies that the EBM practitioner who applies the PICO process is expected to simplify the description of the patient so it make possible a match with the kind of generic descriptions commonly used to identify research populations. In their book "painless EBM" Antonio Dans and colleagues (2008) provide a nice illustration of this. These authors prescribe that the patient's description should not go beyond certain *"disease"* or *"diagnosis"*, for example, *"acute chest pain"* or "HIV" (p. 8). The logic underlying this advice is simple enough: the EBM practitioner should focus on what matters. Why raise a question for which there is no good answer? Is it not better to ask questions that have been investigated by valid methods?

However, as we saw in the first chapter, this kind of selective attention is not harmless, for by recommending EBM practitioners to classify their patients into the broad reference class for which it is likely that valid research will be found, the 4S model, via the PICO process, starts paving the way for a systematic disregard of potentially valuable information about the patient for which there is no good research data. [35]

Of course, it is true that framing clinical questions in a way that facilitates matches of relevance with research evidence does not necessarily lead to recommendations that neglect the PEI. After all, the inferential process is just beginning and it might be that the extra information available about the patient might be considered later on, before arriving at the final recommendations. Nonetheless, if the 4S model truly aimed to practice EBM in accordance with Sackett's definition, then one would expect that this teaching approach

[35] It is not necessary to dwell on this point, but notice that there is a way in which each of the elements of the PICO process (patients, interventions, comparisons, and outcomes) promotes abstractions aimed at matching individuals with the reference classes for which there is valid research data available.

would facilitate attention to all available information from the outset. After all, if Sackett's definition were descriptively accurate, then EBM practitioners would integrate research evidence with information from their own experience. Given this, it seems odd that the 4S model promotes abstractions exclusively directed at facilitating the acquisition of research evidence during the next steps.

So, as we have seen, the first step of the 4S model, via the PICO process, encourages physicians to frame the patient's case so to make the identification of clinical trials much easier and, by contrast, omits any abstractions that could make the search for non-research evidence, or the application of clinical judgment, any less difficult. Thus, up to now, it appears that the recommendations supported by this model are more consistent with my reconstruction than with EBM's official definition. Let us now, however, turn to the second step of the 4S model, for perhaps now, things might get a little better for supporters of EBM.

2.4.1.2. Step 2: Acquiring evidence

The second step of the 4S model advises EBM practioners on how to obtain evidence for clinical decision-making. For this step be aligned with the official definition of EBM, one would expect that it would help physicians to acquire whatever kind of evidence is necessary to make sense of what they know about their patients. However, this second step makes even clearer that the emphasis of the 4S model has been placed in providing EBM practitioners with tools that restrict their attention to clinical trials, and simultaneously overlook the PEI.

Continuing with a narrow focus on research evidence, this step omits advice on how to acquire evidence from sources other than clinical trials. In fact, in their eagerness to facilitate the acquisition of findings from certain types of research, some editions of the 4S model match questions related to traditional medical tasks (e.g. diagnostic, prognostic, and therapeutic questions) with particular research designs, thus suggesting that paying attention to other types of evidence would be a waste of time. For questions of therapy, for example, EBM practitioners have been advised to focus on evidence from RCTs (Sackett et al. 1985), and in the case of diagnostic or prognostic tasks, physicians have been recommended to look for articles on risk models and predictive rules (Adams and Levenson, 2012). And while the link between traditional clinical tasks and particular research designs has been relativized in further editions of the 4S model (e.g. Sackett and Wennberg, 1997), it is remarkable that the methods aimed at facilitating acquisition of evidence included in the 4S model do not include any practical advice as to how to acquire information from what is considered low-ranked evidential sources, such as basic research about mechanisms (Straus et al. 2009; Clarke et al. 2013; Fuller and Flores, 2015).

Examples illustrating that the search for evidence in the 4S model is actually a search for particular clinical trials abound: consider the *PubMed Clinical Queries Feature*, a search tool supported by EBM (Bartkowiak, 2005), which automatically eliminates non-randomized studies for questions of therapy. As the website of the Arizona Health Science Library explains: *"The default when the clinical studies search is set to "therapy" will limit the search to Randomized Controlled Trials"* (The University of Arizona, no date). Or, for another example, take notice of the advice provided by Randolph and colleagues (1999) in one of a series of articles on EBM practice edited by the Journal of the American Medical Association (JAMA). When describing the stage when evidence should be acquired their advice is: *"…you connect to the Internet and decide to search the medical literature…. You limit your search to randomized controlled trials… You choose a Randomized Trial of Corollary Orders…"* (p.67).

As these quotes illustrate, in the context of therapy the search strategies promoted to acquire evidence in the 4S model give preeminent attention to research designs ranked higher in hierarchies of evidence (e.g. RCTs), so a fortiori, this step of the 4S model seems to neglect non-research sources.

Of course, this does not mean that EBM practioners following this model will automatically ignore additional information about their patients or will put their judgment aside. But still it is clear that to encourage physicians to narrow down their attention to research evidence does not count as facilitating attention to whatever information is present in the patient. Obviously physicians might well look for non-research evidence to determine whether the extra features present in their patients are probabilistically relevant to the outcome of interest, but if this happens, it would not be a consequence of the 4S model to train EBM practitioners. So, at the very least, this second step of the 4S model does not make it easier for physicians to address the PEI, and therefore perpetuates the strategic simplifications made in the previous step, which aimed exclusively at finding a match between the best research evidence available and particular individuals.

So, it appears that the recommendations based on the second step of the 4S model remain much more consistent with my reconstruction of EBM as a practice deprived of judgment than with Sackett's portrait of EBM. For, as we have seen, the 4S model seems to be much more concerned with facilitating attention to certain types of valid research than with correcting a potential lack of attention to part of what the physician knows about the patient. Thus, rather than promoting a judicious and sensitive search for whatever kind of evidence useful to estimate the outcome in the light of what the physician knows about the patient, this

second step of the 4S model endorses searches for selected research designs based on a carefully curated summary of what is "known" about the patient.

2.4.1.3. Step 3: Appraising evidence

The divergence between the kind of recommendations endorsed by the 4S model and those that we would expect from EBM practitioners according to Sackett's definition becomes even more apparent at this stage. For, if this model really teaches physicians to practice along the lines of the definition of EBM, one would expect them to acquire tools to appraise the quality of both research and non-research evidential sources. However, instead of providing physicians with a wide range of tools to assess the quality of various kinds of evidential sources, so that they can exercise their judgment regardless of the type of information they happen to have available, this third step of the 4S model ensures that only research evidence ranked at (or towards) the top of hierarchies of evidence be considered as input for recommendations.

This third step is operationalized as what is known as "*the critical appraisal exercise"* (Guyatt, 1991; Green, 1999). In essence, this exercise consists of an assessment of the quality of the evidence available, which is oriented to filter out evidential sources that fail to meet certain quality standards (Henegan and Badenoch, 2002).

In order to understand how the critical appraisal exercise promotes the exclusion of certain types of evidential sources we need to pay attention to what prominent authors in the EBM camp call the first fundamental principle of EBM: the EBM rules or hierarchies of evidence (Montori 2008, p.1815; quoted from Bluhm and Borgerson, 2011).[36]

Note that hierarchies of evidence have shaped EBM practice through different routes [37]. However, their influence in recommendation is particularly apparent in the third step of the 4S model. In this context, EBM rules of evidence are important because evidence rankings are often presented vaguely as rankings that arrange evidential sources with respect to their validity (e.g. La Caze, 2009), quality (e.g. Brozec et al. 2009), and even their clinical usefulness (e.g. Andrews et al. 2013).

In the context of the 4S model, rules of evidence are presented to EBM practitioners as tools to help them select sources with a greater number of epistemological strengths and fewer

[36] In his textbook on EBM, Jeremy Howick (2011) claims that the philosophy of EBM *"is best expressed in EBM hierarchies"* (p.4).
[37] For example, hierarchies of evidence also take part in the construction of recommendations in the EBM guidelines model of practice (§ 2.4.2).

epistemological weaknesses. However, in practice, the quality markers used to rank evidential sources varies from one hierarchy to the next (Blunt, 2015).

So, for example, in hierarchies of evidence designed to guide therapeutic recommendations, well-conducted RCTs are typically placed at the top (or near the top after meta-analyses of RCTs) mainly, but not exclusively, because of their ability to identify causal relationships (Cartwright, 2011; Papineau, 1994). In the context of diagnosis and prognosis, on the other hand, well-conducted prospective studies (for instance, cohort studies) occupy the top position due, among other things, to their capacity to obtain precise and unbiased correlations (Burns et al. 2011). It must be noticed, however, that most hierarchies supported by advocates of EBM (e.g. NICE, 2004; U.S. Preventive Services Task Force, 2008; Canadian Task Force, 1979; Sackett, 1986,1993) seem to identify susceptibility to bias as one of the most important epistemological weaknesses of an evidential source (Blunt, 2015). And this is one of the central reasons why randomized clinical research is typically regarded as epistemically superior to observational studies, basic research on mechanisms and information from clinical experience.[38]

But let me return now to the main point of this section. How are the rules of EBM used during the critical appraisal exercise? As Jacob Stegenga points out, the standard prescription implied in hierarchies of evidence is that physicians ought to take *instances* of evidential sources (evidence tokens) ranked higher in hierarchies to be more reliable than instances of evidential sources placed lower (Stegenga, 2014). And because the place of any given source of evidence is often seen as closely connected to judgments about the clinical usefulness of such evidence (e.g. Belsey, 2009), one of the practical upshots of hierarchies is that the EBM practitioner should spend most of her scarce time paying attention to evidential sources that are ranked high, precisely because these are less likely to mislead and more likely to inform (Sackett et al. 1996).[39]

Let me offer an illustration of how EBM practitioners are advised to exercise the critical appraisal in practice. The *Evidence-based tool-kit* tells physicians that: *"the most important type of research for answering therapy questions is the randomised controlled trial (RCT)."* And then continues to say that if the answer to the question *"were the groups randomized?"*

[38] Notice that although some supporters of EBM have suggested that some aspects of hierarchical rules of evidence, for instance the superiority of RCTs over observational studies and basic research on mechanisms are well-supported by empirical evidence (§ 1.5.2.2), others –for example, Howick (2011)– have, at some moments, recognised that "*the EBM hierarchy itself appears to be supported by "weak" (according to EBM) evidence, namely the opinion of EBM experts!"* (p.6).

[39] As Gordon Guyatt and colleagues (2000) explain: *"The hierarchy implies a clear course of action for physicians addressing patients problems–they should look for the highest available evidence from the hierarchy."* (p.1293).

is *"not satisfactory",* then physicians *"should consider ditching the paper and looking elsewhere"* (Henegan and Badenoch, 2002, p.1). Thus, for questions of therapy, the third step of the 4S model explicitly encourages practitioners to jettison study-types not ranked high in hierarchies of evidence.

Furthermore, the stage at which evidence is appraised in the 4S model also encourages recommendations vulnerable to the PEI through a subtler mechanism. For, as I said at the beginning of this subsection, if the 4S model truly endorses recommendations consistent with Sackett's definition, then one would expect at least some advice (even very general) as to how to assess the quality of non-research evidential sources. After all, if, according to EBM's official definition, EBM practitioners integrate the best research with clinical experience, then an approach to teach that kind of EBM practice should help physicians to distinguish when clinical experience and their judgment can be useful and when they can be misleading. So, while the 4S model, via the critical appraisal exercise, teaches EBM practitioners about various methodological aspects of clinical trials, this model has remained largely silent as to how to assess the reliability of the so-called low-quality evidential sources. Surely, not all instances of judgment informed by clinical experience (or perhaps reasoning based on mechanisms supported by basic research) are equally unreliable or faulty. Won't the specific kind of reasoning applied, as well as the accuracy, completeness, and reliability of the primary evidence involved, make the resulting judgments be more or less reasonable? And yet, despite this, it is remarkable that the third step of the 4S model, which is supposed to help physicians to appraise evidence, only provides physicians with a methodological checklist for selected research designs and ignores the relevant task of assessing the quality and reliability of information from clinical experience.

Of course, this might be an innocent oversight. But regardless of the reasons behind it, the fact that the 4S model does not teach EBM practitioners how to assess the quality of non-research sources seems difficult to reconcile with the idea these practitioners are being trained to put into effect the kind of EBM described in Sackett's definition. As far as I can see, this kind of training prepares physicians to find and select the best research evidence so that it can then be used to back up recommendations for individuals. And if this is correct, it seems like, after its third step, the 4S model keeps helping physicians to focus on selected research evidence. And, in doing so, it continues to foster recommendations that leave little room for judgment and are vulnerable to the PEI.

2.4.1.4. Step 4: Applying evidence

At this stage, advocates of EBM might admit that the previous three steps of the 4S model promote recommendations exclusively based on the information for which valid research data

is likely to be found. However, it might still be too hasty to conclude that the 4S model teaches physicians to practice medicine in a way that is consistent with my reconstruction of the EBM approach and, by contrast, difficult to reconcile with Sackett's version. Such conclusion would be precipitate because supporters of EBM might still contend that it is during the fourth step of the 4S model, the step where valid evidence is applied to individuals, that EBM practitioners exercise their judgment and address the PEI.

I accept that there is some plausibility to this defence. If it is true that in the fourth step of the 4S model physicians are entrusted with the task of determining the "applicability" of the evidence gathered in the previous steps, and if "applicability" is defined at least in some EBM textbooks –for example the one written by Antonio Dans and colleagues (2008)– as *"the extent to which conclusions of a study can be expected to hold true for a particular patient"* (p.11), one might think that this model does not seem to be oblivious of the PEI. And if so, by encouraging judgments of applicability, the 4S model might be teaching EBM practitioners to use their clinical judgment and derivatively encourage them to consider information beyond valid research findings.

However, if supporters of EBM were right on this, the fourth step of the 4S model would have to offer a significant compensation for the neglect of additional information inherited from the previous steps. One way to do this would be to foster explicit awareness about the importance of the PEI so to encourage EBM practitioners to apply their judgment in the light of everything they know about their patient, regardless of whether the relevance of the information available could only be supported by low-ranked evidential sources.

But I think that this would be to ask too much of the 4S model. Below I shall argue that, in practice, the influence of EBM rules of evidence prevails. Even if there are some timid calls for judgment when it comes to applying research findings, it is unlikely that EBM practitioners will overcome the pressure to generate recommendations that, in accordance with EBM rules, must be based on evidence of the highest quality available.

Thus, despite first appearances, I shall argue that the 4S model ends discouraging EBM practitioners from paying attention to information about the patient whose relevance can only be defended by appealing to clinical judgment itself (e.g. in the form of experts' opinion) or by citing basic research or previous personal clinical experience with similar patients. If explicitly defending the incorporation of such information in clinical recommendations is difficult enough in the absence of supporting research evidence, it is almost impossible when EBM practitioners know that there is valid research evidence for a target population to which the individual patient belongs to supporting another recommendation. So, if clinical discretion were to be exercised by the end of the 4S model, it would only be exercised in a very

restricted sense – perhaps only under extreme circumstances, so to prevent blatant misapplications of research data to individuals.

2.4.1.4.1. "Non-exclusion" as applicability

Let us now consider concrete guidance with regard to assessments of applicability in the context of the 4S model. A first problem for those who claim that in this last step the 4S model can correct the neglect of available information fostered in previous steps is that some supporters of EBM provide physicians with advice that is, quite simply, misguided. I say this because the advice offered, rather than helping EBM practitioners to widen their gaze so to look for any kind of evidence they might need to make sense of the information present in the patient and estimate the right probabilities, avoids addressing the PEI by asking the wrong questions.

Examples of this kind of "misguidance" can be found in the writings of EBM authors who seem to think that the question of the applicability to individuals is interchangeable with the question of whether the patient could have been included in the study sample. An error, which as the reader might recall from the first chapter, confuses the problem of determining what target population research evidence is representative of (that is, the problem of generalizability or external validity) with the PEI, which consist of knowing something extra about the particular individual the physician is interested in. Consider, as illustrations, the following quotes:

"Could my patient have been randomised in this trial? If so the results are applicable; if not, they may not be."

(Valori, 2002, quoted from Summerskill, 2005 p.13)

Or, even more explicitly,

"…how does the cardiologist know whether to apply a specific treatment--whose efficacy and risk profile have been defined in a good quality RCT… to his/her individual patient? The simplest solution would be found in the answer to the question: "Could my patient have been enrolled in the study by satisfying its inclusion and exclusion criteria?" If the answer is "yes", then it would be sound to apply the study results to the patient…" (Carneiro, 2003, p.259)[40]

[40] Similar quotes are "*Could my patient have been included in this study? (Does my patient meet the inclusion and exclusion criteria used in this study?)"* (Nordenstrom, 2007, p. 69) Or: "*The applicability of clinical trial results to the individual patient depends on a rigorous set of rules that can be summarized in the question "Could my patient have been enrolled in this trial?"* (Carneiro, 2003, p.259). See also Sackett, 2000 for analogous counsels.

As the previous passages suggest, EBM practitioners who follow this advice during the last step of the 4S model are taught that the question of the applicability of research to individuals is satisfactoryly solved when patients could have been included in the study sample.

On a charitable reading, one could see that this approach to assess applicability is well intended. I say this because at least it raises the possibility that the results of certain clinical trials, however valid, might not be applicable to some individuals. But having said that, it is apparent that this kind of advice does not grasp the crux of the PEI, and therefore does not address it properly. To see this, one only needs to point out that paying attention to inclusion and exclusion criteria may, sometimes, be equivalent to paying attention to everything the physician knows about the patient; but given the way in which patients' data are normally collected by physicians in real clinical settings (Drill et al. 2015, Dugdale et al. 1999; Walker et al. 1990; Kaplan et al. 1997), it is reasonable to assume they will know more about their patients than whether they meet or do not meet a few inclusion and exclusion criteria[41]. Thus, this kind of advice is unlikely to redirect physicians' attention towards the patient's idiosyncrasies that were extracted away during the previous steps of the 4S model, and therefore, is likely to perpetuate recommendations that forcibly eliminate the PEI and encourage the application of valid research findings on the basis of membership of standard target populations.

As I pointed out in the previous chapter, if the PEI is taken seriously, then it follows that the main challenge for the physician is to use everything she knows about the patient to estimate what the relevant probabilities are, and it is quite clear that this challenge may still arise after the physician has no doubts about the fact that, given certain inclusion and exclusion criteria, her patient could have been included in the study sample.

Finally, it worth stressing that this kind of approach to implement the fourth step of the 4S model does little or nothing to align the recommendations of EBM practioners who follow this criteria to those ascribed to EBM practice by Sackett's definition. For to encourage physicians to determine whether inclusion or exclusion criteria would leave the patient out of the trial is undoubtedly different from encouraging physicians to integrate the best research evidence with "their judgment informed by experience". So, if this is all there is to the assessment of applicability in the 4S model, it would be difficult to claim that it will turn resulting recommendations congruent with EBM's official definition and inconsistent with my representation of EBM

2.4.1.4.2. Applicability by default

[41] To complicate things even more, it is well known that inclusion and exclusion criteria are poorly reported by many researchers (See Blümle et al., 2011).

However, the inclusion/exclusion criteria approach is not the only alternative, for another way to operationalize judgments of applicability in the context of the 4S model has been proposed by Matt Petticrew and Sir Ian Chalmers (2011) [42].

In essence, Petticrew and Chalmers' approach to applicability in the context of the 4S model is that valid clinical trials (in particular RCTs) should be assumed applicable to every member of the intended target population except in the presence of strong reasons to think otherwise. In response to Nancy Cartwright's remarks on the necessity of a reasoned approach to examine the generalizability of the results from RCTS (Cartwright, 2011), these authors proposed that physicians should approach to the question of applicability by asking themselves: *"Are there any good reasons to believe that the research is not relevant to us, that 'it won't work for us'?"* (p.1696). So far so good, no surprise there, however, Petticrew and Chalmers (2011) continued to put forward the following concrete advice: *"a good working assumption is that the main result probably applies to everyone, unless good evidence exists to the contrary"* (p.1696).

If we focus on the first part of Petticrew and Chalmers' advice, it may appear that EBM practitioners are being taught to incorporate previously neglected information about the patient. To the extent that "reasons" are requested before applying research finding, one cannot say that the 4S model is telling physicians to extrapolate findings automatically on the basis of membership to the target population as my reconstruction suggests. Furthermore, to the extent that these authors' request for reasons involves attention to what physicians know about their patient, then it seems that Petticrew and Chalmers' counsel is better suited to address the PEI. After all, for the request for potential reasons not to apply research findings to individuals to have any practical influence in final recommendations, these authors should leave some room for the exercise of clinical judgment.

Nonetheless, attention to the second part of Petticrew and Chalmers' recommendation (2011), what they called *"a good working assumption"* (p.1696), reveals that these authors impose important constraints on the application of clinical judgment, and do not promote attention to the PEI. For according to this assumption, the claim that valid research findings are applicable can only be challenged under special circumstances, namely: in the presence of *"good evidence"* (p.1696).

The demand for "good evidence", although in principle sensible, becomes problematic because EBM practitioners the notion of good evidence is tightly connected to EBM rules,

[42] Sir Ian Chalmers is a very influential figure in the area of EBM. He has not only written hundreds of articles supporting reliance on RCTs, but also was the founder of the Cochrane collaboration, which aims at providing physicians with updated summaries of the best evidence available.

and therefore these rules affect judgments as to what constitute good reasons not to extrapolate evidence. So, from this perspective, EBM practitioners are perfectly entitled to establish the applicability of research without attending to everything they know about the patient, and, what is more, are in effect encouraged to determine applicability by restricting their attention to information whose relevance is defensible according to EBM rules.

By now, the reader will have noted that Petticrew and Chalmers' advice has taken EBM practitioners round a small circle. Because what Petticrew and Chalmers' are actually saying to EBM practitioners is that the reasons not to apply RCT evidence should be backed up by information about the patient whose relevance to the matter has been itself previously backed up by suitably controlled research! Or, to put it the other way around, that EBM practitioners should not waste their time paying attention to information about the patient that is not backed up by research evidence, for such information cannot possibly challenge the "reasonable working assumption" that research results are by default applicable to individuals. In effect, therefore, the recommendations that follow from Petticrew and Chalmers' advice are perfectly consistent with my reconstruction of EBM recommendations as exclusively based on valid research evidence, and which assume that any extra information about the patient not backed up by suitable research is probabilistically irrelevant to the outcome of interest.

So, to summarise, the last step of the 4S model seems of little value against the PEI. And the main reason for this is that it seems implausible that a significant proportion of what the physician knows about her patient could be defended as relevant by citing previously conducted controlled trials. In the context of therapy, for example, that would imply that each potential interaction effect should gain such status on the basis of well-conducted RCTs, which is in practice unfeasible [43].

Furthermore, with respect to the idea that the last step of the 4S model would permit integration with low-ranked evidential sources so as to generate recommendations consistent with Sackett's definition of EBM, it seems highly implausible that this happens in a context where EBM rules dictate the attitude of EBM practitioners to different evidential sources. Suppose, for the sake of emphasis, that a physician has a patient in front of her and she recalls from her experience that she has seen many patients of this kind during her years of practice. Suppose further that, according to the physicians' experience, patients like the patient in question are not likely to respond to the treatment recommended by a well-conducted RCT, whose intended target population, according to the authors of this trial, are patients with the

[43] Notice that since most of our knowledge about interaction effects does not comes from main estimates reported in clinical trials many supporters of EBM would consider such knowledge too unreliable to be taken seriously (Sleigh, 2010; Senn and Harrell, 1997; Burke et al., 2015).

condition this patient has. Now, the question is: should the physician retrace her own steps, abandon the abstractions that made possible attention to high-quality evidence and now appeal to the presence of extra information in her patient so to challenge the applicability of such evidence to her patient? According Sackett's definition, such kind of relevant experience should not be ignored, and what is more, it should be combined with the findings that come from research. But, would the EBM practitioner (in particular a dutiful one) be willing to do that? Surely, as some advocates of EBM have said, if the physician is going to disregard RCT evidence and rely on her judgment she has to be sure of what she is doing (Howick, 2011). But the interesting thing here is that it is difficult to imagine that any physician who adheres to EBM rules of evidence could possibly be willing to jettison RCT findings that superficially seem perfectly applicable so to be able to face the PEI and use her judgment to estimate the right probabilities on the basis of all information available.

In conclusion, the EBM practioners who follows the 4S model faces an evident predicament. While it seems, in theory, possible to combine evidence from various sources when assessing applicability, it is difficult to see how this can be done without violating EBM rules of evidence. Thus, if the applicability of valid research to the individual can only be questioned by information about the individual that is also supported by "valid evidence", we are forced to conclude that the 4S model of EBM practice fosters recommendations almost exclusively concerned with the best research, which leave little room for clinical judgment and which remain vulnerable to the problem of extra-information. As Sandra Tanenbaum (2012) puts it: *"For EBM, clinical science is so far superior to any other form of medical knowledge that there are few good reasons not to act on it, even in the individual case."* (p.272).

2.4.2. Evidence-based guidelines model

The second method through which EBM practice has been implemented is centred in the construction and application of evidence-based guidelines (EBGs). Unlike the 4S model, the practice of EBM via EBGs does not rely on physicians' capacity to search for and subsequently apply high-quality research evidence to their patients. In the EBG model, physicians are presented with clinical recommendations backed-up by the best research evidence available, which has been searched, selected, and analysed by guideline-makers (Lim et al. 2008; Woolf et al. 1999).

Given that EBGs are tools primarily constructed for purposes of public health, one might wonder: what should the practice that follows from EBGs look like for it to be compatible with Sackett and colleagues' integrative definition? Obviously, the fact that guideline-makers develop recommendations for standard reference classes of patients makes it unreasonable to demand attention to knowledge about particular patients during the process of construction of

EBGs. However, this does not imply that a system of care based on EBGs is necessarily incompatible with Sackett and colleagues' definition. Even if recommendations from EBGs are exclusively based on research evidence, their application to individuals can, in principle, coexist with the exercise of clinical judgment and thereby permit attention to the idiosyncrasies of particular patients. Thus, one might think that physicians could interpret EBGs as general suggestions based on simple rules, which could be useful for some patients, but which also could be put to one side when they have access to additional information that permits more sophisticated decision algorithms.

As it happens, the very definition of EBGs seems to support the idea that their application is consistent with Sackett and colleagues' definition, for EBGs are officially presented to physicians as recommendations not mandates. The Institute of Medicine in the United States (Field and Lohr, 1990), for example, emphasises the non-compulsory character of EBGs by defining them as *"systematically developed statements to assist practitioner decisions about appropriate healthcare for specific clinical circumstances"* (p.38). Likewise, the American college of Physicians (ACP) tells physicians that the goal of ACP guidelines is to *"provide clinicians with recommendations based on the best available evidence; to inform clinicians of when there is no evidence; and finally, to help clinicians deliver the best health care possible."* (ACP, no date) And, similarly, in the United Kingdom, several guidelines generated by the National Institute for Health and Care Excellence (NICE) present part of their content using an advisory tone. For instance, the NICE guideline [CG90] for the management of depression in adults specifies that physicians are expected to exercise their judgment because recommendations are *"not mandatory"(*NICE, 2016).

Thus, if EBGs truly are suggestions not commands, then surely their application must be flexible enough to leave sufficient room for the exercise of judgment and thereby avoid the PEI. And if so, the thesis that I mounted a straw-man against EBM would find support in the EBGs model of practice.

Nonetheless, the way in which EBGs are officially presented –either in general definitions or in specific guidelines– does not provide a definite answer to the question of what kind of clinical practice follows from the application of EBGs. Caution is necessary here, for although standard definitions and the presence of provisos in some EBGs are important, these are only part of a larger set of mechanisms through which EBGs shape physicians' clinical behaviour.

In fact, I think that it would be rather naïve –if not conveniently simplistic– to assume that careful definitions and statements with qualified conditions of use determine the way in which EBGs are in practice applied. As several authors have noted, the relationship between

EBGs and clinical practice is more complex than it appears, and to fully understand how EBGs really shape the care of individuals we need to consider how EBGs relate to standards of practice and measures of quality and performance (e.g. Bogdan-Lovis et al. 2012, Tanenbaum, 2012). This is because physicians' compliance (or non-compliance) with standards and quality measures brings about professional, economical and legal consequences, which in turn affect how physicians interpret the compulsoriness of EBGs in the first place (Hayward and Kent 2008, Shackelton et al. 2009, Karve et al. 2008). Below I shall direct the reader's attention to a set of implicit and explicit influences, which jointly counteract the advisory character of EBGs, and end by promoting a rigid kind of EBM practice with an exiguous space for clinical discretion.

2.4.2.1. The tacit influence of EBM rules of evidence

One of the reasons that makes it difficult for physicians not to comply with EBGs, even if they are presented as recommendations, and even if physicians are requested to exercise judgment, was already discussed in the context of the 4S model: the pressure to practice in accordance with the rules of EBM.

When it comes to deciding whether to apply a recommendation for a particular patient it is hard to see how a physician who wishes to be identified as an evidence-based practitioner would dare to challenge recommendations supported by *"the best scientific evidence available"* (Grol et al. 1998. p.858).

The influence of EBM's rules on physicians' attitude towards EBGs is illustrated vividly in the comments of Peter Szatmari (2004), a physician who expresses his frustration in response to the problem of basing recommendations on evidence of dubious applicability to individuals (Gupta, 2004): *"On what basis do I make a clinical decision? … I would prefer to make my clinical decision on the basis of the best available evidence. What is the alternative to using evidence as a guide? I could do nothing…I could make a decision at random… I could make my decision on the basis on how I was trained…[or] I could instead make my decision based on what I know about pathophysiology…Do we have an alternative?"*(p.97-98).

Many physicians who, like Szatmari, adhere to the rules of EBM, feel that when EBGs are presented as supported by valid research evidence, there seems to be no way in which they could challenge recommendations. This idea was expressed when I introduce the case of Mr Smith in the previous chapter: If the physician knows that her patient has acute conjunctivitis, and she has access to valid EBGs for such patients: why bother in exercising her judgment to pay attention to extra information whose probabilistic relevance to the outcome of interest is difficult to defend according to EBM rules?

The point is simple enough, if attention to additional information might lead to recommendations that diverge from EBGs, and which could be criticised on the grounds that the relevance of such information can only be backed up by low-quality evidence, then the systematic exercise of clinical judgment to determine the applicability of EBGs seems to be futile. And, in consequence, it would not be surprising if most practical-minded physicians, who work for institutions that embrace EBM rules, would prefer to ignore the PEI and just apply EBGs on the assumption that the extra facts they might know about their patients are probabilistically irrelevant.

But even if some physicians were sufficiently aware about the negative consequences of ignoring the PEI, and therefore were interested in exercising their judgment during the application of EBGs, there are stronger and more explicit influences than a potential conflict with EBM rules, which might persuade them that it is more convenient to put their clinical judgment to one side.

2.4.2.2. Standards of care and "best practices"

Nowadays there is a clear link between EBGs and standards of clinical practice. In the United Kingdom, for example, many recommendations included in *NICE guidelines* are explicitly tied to *NICE quality standards* (NICE, 2016a). These standards consist of "…*markers of high-quality, cost-effective patient care, covering the treatment and prevention of different diseases and conditions. Derived from the best available evidence such as NICE guidance …"* (NICE, 2016b).

It is worth stressing, however, that the correspondence between EBGs and quality measures is not absolute. In this respect, not all recommendations included in NICE guidance gain the status of quality indicators. This is not surprising, because quality standards are normally established by a process that not only involves attention to research evidence but also additional considerations, such as feasibility and economical costs at a population level (NICE, 2016b). Nonetheless, the fact that the link between EBGs and quality standards is not applicable to all recommendations, does little to change the basic message conveyed by this connexion: while compliance with EBGs suggests that clinical care is of the highest-quality, non-compliance is assumed to be indicative of sub-optimal practice.

The situation is not very different in United States, where the quality of care has been operationalised using performance indicators largely influenced by EBGs (Tannebaum, 2012). Thus, Rodney Hayward and David Kent (2008) two primary care physicians ironically comment: *"providing the highest quality of care is no longer the challenging task it used to be, involving clinical acumen and understanding of individual circumstances, concerns and*

needs of your patients. Those days are thankfully gone, replaced with a far more reasonable principle – providing highest quality care simply means scoring well on your performance measures." (p 255)

Notice, however, that the point I am pursuing here is not that establishing a link between EBGs and practice standards is necessarily a bad thing. I seriously doubt that any reasonable person would reject that a certain level of practice standardization is necessary for a public system of health to work efficiently, and that performance measures can be sensible way to check compliance with established standards. My point is rather that the connexion between EBGs, standards of practice, and quality measures need not be too rigid. If it lacks flexibility, it introduces an additional and potentially detrimental incentive to comply with EBGs. For, if the link with quality measures eliminates clinical discretion, it is plausible to think that physicians (regardless of whether they style themselves as EBM practitioners and are concerned with following EBM rules of evidence) would be concern with complying with EBGs if they want to avoid being considered professionals who practice suboptimal clinical medicine.

But, of course, the connexion between EBGs and quality measures need not result in a model of care that forces physicians to apply EBGs uniformly. One can conceive of a reasonable system of quality measures that leaves enough room for exceptions. In such a system departures from EBGs would not be immediately taken as instances of suboptimal practice, and if so, physicians would feel less compelled to adhere invariably to EBGs. Regrettably, although there is increased awareness as to the importance of generating quality standards that are sensitive to special cases (e.g. Mercuri and Gafni, 2011), the most common scenario seems to be one in which departures from EBGs are very quickly classified as cases of deficient medical care.

As Falzer and Garman (2009) point out, the *"lack of conformance to clinical guidelines and dosage recommendations have been attributed to factors such as lack of knowledge, vulnerability to the persuasive appeals of big pharma, intransigence, poor training, or lack of proper monitoring and inherent limitations of human cognitive functioning."* (p.1143). So, it seems that the standard attitude towards non-compliance with EBGs is not to assume that physicians were able to provide patients with a better recommendations based on more information, but rather that physicians actually failed to prescribe the right interventions (e.g. Rapp et al. 2008; Schoenwal et al. 2008).

Furthermore, even if risking being considered a lazy professional who does not deliver high-quality care to her patients is not sufficient reason to put judgment to one side, there are

additional incentives that might lead the physician to favour an automatic conformance to EBGs, even in the presence of information suggesting that the patient would be better off if she does not comply with EBGs.

2.4.2.3. Pay for performance initiatives

One of the most obvious positive incentives to comply with EBGs is economic, for nowadays the physician who practices in accordance to guideline recommendations can increase her income. McCartney and colleagues (2016) describe the current situation in the United Kingdom:

"[EBGs] have been used to create financial incentive schemes such as the UK's Quality and Outcomes Framework, whereby a substantial proportion of general practice income depends on achieving thresholds for drug therapy or surrogate outcomes in accordance with National Institute for Health and Care Excellence guidelines" (p.1).

As McCartney and colleagues (2016) say, the scope and impact of financial incentives is considerable. In a report on the experience of pay for performance incentives in the United Kingdom, Campbell and colleagues (2009) indicated that *"payments make up approximately 25% of family practitioners' income, and 99.6% of family practitioners participated in the pay-for-performance scheme…"* thus confirming that these stimuli may play an important role in shaping physicians' clinical practice (p.369).

However, although it is reasonable to assume that economic incentives do not affect the behaviour of every physician in the same way (Wieringa and Greenhalgh, 2015; Montgomery, 2006, Elwyn et al. 2016), there is widespread concern that they are turning evidence-based "guidelines" into evidence-based "tramlines". As McCartney and colleagues (2016) say: *"many clinicians and patients have expressed dissatisfaction with the way evidence based medicine has been applied to individuals, especially in primary care. There is concern that guidelines intended to reduce variation and improve the quality of care have instead resulted in medicine becoming authoritarian and bureaucratic"* (p.1). And while most commentators (e.g. Greenhalgh et al. 2014; Harrison and Wood, 2000; McCartney et al. 2016) have pointed out that the use of pay for performance incentives encourages neglecting patients' preferences and values, it is clear that a system that rewards rigid application of EBGs also discourage physicians from paying attention to the additional information they might have about individual patients. The mechanism is simple: If the physician knows that the consequences of exercising her judgment are not only recommendations that might be difficult to defend, but also recommendations that will not bring about any economical benefit, why bother about the PEI? After all, if heath-care managers and supervisors do not worry about the

consequences of ignoring extra information, why the physician should worry about this problem?

Even so, one might hope that not all physicians are willing to put their own economic benefit above their patients' interest. (In fact, as a clinical psychiatrist, I would like to think of myself as one of those physicians.) And in fact, studies that reveal physicians' dissatisfaction with the way in which they are encouraged to apply EBGs suggest that at least some of physicians seem to be more interested in improving their patients' health than in increasing their income (Harrison and Wood, 2000). Nonetheless, there is yet a further incentive to comply with EBGs, which might be even more persuasive for it appeals to a particular kind of fear, namely, fear of malpractice lawsuits.

2.4.2.4. Defensive medicine and risk of litigation

Concerns of malpractice liability are particularly prominent in developed countries such as United Kingdom and United States, where the term "defensive medicine" was first coined (Bishop and Pesko, 2015). This term refers to a situation where physicians' recommendations are primarily aimed at decreasing the risk of litigation (Sekhar and Vyas, 2013).

Of course, defensive medicine is a problem of its own, which is not necessarily related to EBGs. However, to the extent that defensive medicine involves compliance with recommendations that are more defensible in case of adverse outcomes, it is plausible to think that physicians whose recommendations are driven by fear would prefer to comply with EBGs, even in a situation where EBGs do not provide them with the recommendation that maximizes the probability of the outcome of interest for the patient.

Consider the following quote from the Department of Health (1999) taken from a comment on the role of EBGs in determinations of medical negligence by Brian Hurwitz (2004): *"Any doctor not fulfilling the standards and quality of care in the appropriate treatment that are set out in these Clinical Guidelines, will have this taken into account if, for any reason, consideration of their performance in this clinical area is undertaken."* (p.1024).

The point illustrated by this quote is made explicit by Steven Woolf and colleagues (1999): EBGs constitute *"citable evidence for malpractice litigation"* (p.530), and since several studies suggest that fear of malpractice occupies a prominent role among the factors determining practice styles (Schumacher et al. 1995; Carrier et al. 2010; Mercuri and Gafni, 2011), it would not be surprising if physicians interested in protecting themselves against litigation opt to ignore the PEI and just practice in accordance to EBGs.

Of course, it is true that EBGs do not necessarily establish legal standards of practice, and therefore physicians are not "forced" to follow EBGs. However, EBGs do not require a legal status to exert a powerful behavioural influence on physicians' practice. As Hurwitz (2004) points out, EBGs *"do provide the courts with a benchmark by which to judge clinical conduct"*, and this implies that they *"set normative standards such that departure from them may require some explanation"* (p.1028). And while such demand for explanation may not be a problem for some physicians, it is reasonable to assume that most physicians would prefer to avoid formal inquiries into their practice, even if they know that they can argue that given the information available about the patient at the time of decision-making their recommendations were the right ones.

So, to summarize, although the fact that EBGs are normally presented as recommendations seems to support the thesis that they leave enough room for the exercise of clinical judgment, careful attention to the relationship between EBGs and standards of care reveals that EBGs take part in a model of clinical care that is far more rigid than initially appears. In practice, physicians are encouraged to comply with guidelines by at least three mechanisms: first, compliance with EBGs helps them to maintain their professional reputation as EBM practitioners; second, it helps them to increase their income; and third, it may protect them in case of being subject to malpractice lawsuits. Thus, even though a model of practice based on EBGs is in principle compatible with Sackett's definition, in practice it encourages physicians to put their judgment to one side and therefore is more consistent with my reconstruction of EBM.

2.5. Conclusions

In this chapter have I addressed the question of what EBM practice is. My starting point was the so-called EBM integrative definition. By paying attention to the aims stated in this definition I offered a general characterization of the kind of practice that follows from this definition. Then I proposed an alternative description of EBM, as a practice almost completely focused on applying selected research to the care of individuals. Given the incompatibility between EBM's integrative characterization and my characterization I questioned the capacity of EBM's integrative characterization to describe how EBM is actually practiced. I showed that, the practice of EBM that follows from its methods is in tension with the practice of EBM implied by the integrative definition. Finally, I considered the possibility that the inconsistencies between different accounts of EBM practice could be explained by errors in EBM's methods. However, after considering various reasons, I concluded that it is more likely that EBM integrative definition is, in truth, a restrictive kind of EBM practice, in disguise.

2.6. References

Adams, S.T. and Levenson, S.H. (2012). Clinical predictions rules. *BMJ.* 344:d8312.

ACP. [no date]. *"Clinical Guidelines and Recommendations from the American College of Physicians"* Retrieved 05 October, 2015, from https://www.acponline.org/clinical-information/guidelines

Akobeng, A.K. (2005). Principles of evidence based medicine. *Arch Dis Child.* 90:837-40.

Andrews, J.C., Schünemann, H.J., Oxman, A.D., Pottie, K., Meerpohl, J.J., Coello, P.A., Rind, D., Montori, V.M., Brito, J.P., Norris, S., Elbarbary, M., Post, P., Nasser, M., Shukla, V., Jaeschke, R., Brozek, J., Djulbegovic, B. and Guyatt, G. (2013). GRADE guidelines: 15. Going from evidence to recommendation— determinants of a recommendation's direction and strength. *J Clin Epidemiol.* 66(7):726-35.

Bartkowiak, B.A. (2005). Searching for Evidence-Based Medicine in the Literature Part 2: Resources. *Clin Med Res*. 3(1): 39-40.

Belsey, J. (2009). "*What is Evidence-based Medicine?*" 2nd ed. What is…? Series. Available at http://www.medicine.ox.ac.uk/bandolier/painres/download/whatis/ebm.pdf. [Accessed 05 December 2015].

Bishop, T.F. and Pesko, M. (2015). Does defensive medicine protect doctors against malpractice claims? *BMJ.* 351:h5786.

Bluhm, R. and Borgerson, K. (2011). *Evidence-based medicine*. In *Philosophy of medicine: Vol. 16. Handbook of the philosophy of science.* D.M. Gabbay, P. Thagard, ans J. Woods (eds). Amsterdam, Elsevier. pp. 203-238.

Blümle, A., Meerpohl, J.J., Rücker, G., Antes, G., Schumacher, M. and von Elm, E. (2011). Reporting of eligibility criteria of randomised trials: cohort study comparing trial protocols with subsequent articles. *BMJ.* 342:d1828.

Blunt, C.J. (2015). *Hierarchies of evidence in evidence-based medicine.* PhD Thesis, The London School of Economics and Political Science.

Bogdan-Lovis, E., Fleck, L. and Barry, H.C. (2012). It's NOT FAIR! Or is it? The promise and the tyranny of evidence-based performance assessment. *Theor Med Bioeth.* 33(4):293-311.

Brozek, J.L., Akl, E.A., Jaeschke, R., Lang, D.M., Bossuyt, P., Glasziou, P., Helfand, M., Ueffing, E., Alonso-Coello, P., Meerpohl, J., Phillips, B., Horvath, A.R., Bousquet, J., Guyatt, G.H. and Schünemann, H.J.; GRADE Working Group. (2009). Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. *Allergy.* 64(8):1109-16.

Burke, J.F., Sussman, J.B., Kent, D.M. and Hayward, R.A. (2015). Three simple rules to ensure reasonably credible subgroup analyses. *BMJ.* 351:h5651.

Burns, P.B., Rohrich, R.J. and Chung, K.C. (2011). The levels of evidence and their role in evidence-based medicine. *Plast Reconstr Surg.* 128(1):305-10.

Campbell, S.M., Reeves, D., Kontopantelis, E., Sibbald, B. and Roland, M. (2009). Effects of pay for performance on the quality of primary care in England. *N Engl J Med*. 361(4):368-78.

Canadian Task Force on the Periodic Health Examination. (1979). The Periodic Health Examination. *Can Med Assoc J.* 121(9):1193-254.

Carneiro, A.V. (2003). Applicability of Clinical Trial Results to the Individual Patient: Pratical Guidelines. *Rev Port Cardiol.* 22(2):259-68.

Carrier, E.R., Reschovsky, J.D., Mello, M.M., Mayrell, R.C. and Katz, D. (2010). Physicians' fears of malpractice lawsuits are not assuaged by tort reforms. *Health Aff (Millwood).* 29(9):1585-92.

Cartwright, N. (2011). A philosopher's view of the long road from RCTs to effectiveness. *Lancet.* 377(9775):1400-1.

Clarke, B., Gillies, D., Illari, P., Russo, F. and Williamson, J. (2013). The evidence that evidence-based medicine omits. *Prev Med.* 57(6):745-7.

Djulbegovic, B., Guyatt, G.H. and Ashcroft, R.E. (2009). Epistemologic inquiries in evidence-based medicine. *Cancer Control.* 16(2):158-68.

Dans, A.L., Dans, L.F. and Silvestre, M.A.A. (2008). *Painless Evidence-Based Medicine.* San Francisco, John Wiley & Sons Ltd.

Drill, R., Nakash, O., DeFife, J.A. and Westen, D. (2015). Assessment of clinical information: Comparison of the validity of a Structured Clinical Interview (the SCID) and the Clinical Diagnostic Interview. *J Nerv Ment Dis.* 203(6):459-62.

Dugdale, D.C., Epstein, R. and Pantilat, S.Z. (1999). Time and the Patient–Physician Relationship. *J Gen Intern Med.* 14(Suppl 1): S34-40.

Eddy, D.M. (1990a). The Challenge. *JAMA.* 263(2):287-90.

Eddy, D.M. (1990b). Clinical decision making: from theory to practice. Practice policies --what are they? *JAMA.* 263(6):877-8, 880.

Eddy, D.M. (1990c). Practice Policies: Where Do They Come From? *JAMA.* 263(9):1265, 1269, 1272.

Eddy, D.M. (1990d). Clinical decision making: from theory to practice. Practice policies--guidelines for methods. *JAMA.* 263(13):1839-41.

Eddy, D.M. (1990e). Clinical decision making: from theory to practice. Guidelines for policy statements: the explicit approach. *JAMA.* 263(16):2239-40, 2243.

Eddy, D.M. (1990f). Clinical decision making: from theory to practice. Designing a practice policy. Standards, guidelines, and options. *JAMA.* 263(22):3077, 3081, 3084.

Elwyn, G., Wieringa, S., and Greenhalgh, T. (2016). Clinical encounters in the post-guidelines era. *BMJ*. 353:i3200.

Falzer, P.R. and Garman, M.D. (2009). A conditional model of evidence-based decision making. *J Eval Clin Pract.* 15(6):1142-51.

Field, M.J. and Lohr, K.N. (1990). *Clinical Practice Guidelines: Directions for a New Program.* Washington, National Academies Press.

Flodgren, G., Hall, A.M., Goulding, L., Eccles, M.P., Grimshaw, J.M., Leng, G.C. and Shepperd, S. (2016). Tools developed and disseminated by guideline producers to promote the uptake of their guidelines. *Cochrane Database Syst Rev*. 8:CD010669.

Fuller, J. and Flores, L.J. (2015). The Risk GP Model: the standard model of prediction in medicine. *Stud Hist Philos Biol Biomed Sci.*54:49-61.

Goldacre, B. (2006). *Objectionable 'objectives'* [Online]. The Guardian. Available at https://www.theguardian.com/science/2006/aug/19/badscience.uknews [Accessed 05 December 2015].

Green, M.L. (1999). Graduate medical education training in clinical epidemiology, critical appraisal, and evidence-based medicine: a critical review of curricula. *Acad Med.* 74(6):686-94.

Greenhalgh, T., Howick, J. and Maskrey, N.; Evidence Based Medicine Renaissance Group. (2014). Evidence based medicine: a movement in crisis? *BMJ.* 348:g3725.

Grimshaw, J.M. and Russell, I.T. (1993). Achieving health gain through clinical guidelines. I: Developing scientifically valid guidelines. *Qual Health Care.* 2(4):243-8

Grimshaw, J., Eccles, M., Thomas, R., MacLennan, G., Ramsay, C., Fraser, C. and Vale, L. (2006). Toward evidence-based quality improvement. Evidence (and its limitations) of the effectiveness of guideline dissemination and implementation strategies 1966-1998. *J Gen Intern Med.* 21(Suppl 2):S14-20.

Grol, R., Dalhuijsen, J., Thomas, S., in't Veld, C., Rutten, G. and Mokkink, H. (1998). Attributes of clinical guidelines that influence use of guidelines in general practice: Observational study. *BMJ.* 317:858–61.

Grol, R. and Grimshaw, J. (2003). From best evidence to best practice: effective implementation of change in patients' care. *Lancet.* 362(9391):1225-30.

Gupta, M. (2004). Evidence-based medicine: ethically obligatory or ethically suspect? *Evid Based Ment Health.* 7(4):96-7.

Guyatt, G.H. (1991). Evidence-based medicine. *ACP J Club.* 114:A16.

Guyatt, G.H., Haynes, R.B., Jaeschke, R.Z., Cook, D.J., Green, L., Naylor, C.D., Wilson, M.C. and Richardson, W.S. (2000). Users' Guides to the Medical Literature. XXV. Evidence-Based Medicine: Principles for Applying the Users' Guides to Patient Care. *JAMA.* 284(10):1290-6.

Harrison, S. and Wood, B. (2000). Scientific-bureaucratic medicine and UK health policy. *Review of Policy Research.* 17(4):25-42.

Haynes, R.B. (2006). Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based health care decisions. *ACP J Club*. 145(3):A8.

Hayward, R.A., and Kent, D.M. (2008). 6 EZ steps to improving your performance: (or how to make P4P pay 4U!). JAMA. 300(3):255-6.

Henegan, C. and Badenoch, D. (2002). *Evidence-based Medicine Toolkit*. London, BMJ Books.

Howitz, J. (2011). *The Philosophy of Evidence-based Medicine.* Oxford, Wiley-Blackwell.

Hurwitz, B. (2004) How does evidence based guidance influence determinations of medical negligence? *BMJ.* 329(7473):1024-8.

Kaplan, C.B., Siegel, B., Madill, J.M. and Epstein, R.M. (1997). Communication and the medical interview. Strategies for learning and teaching. *J Gen Intern Med*. 12(Suppl 2):S49-55.

Karve, A.M., Ou, F.S., Lytle, B.L. and Peterson, E.D. (2008). Potential unintended financial consequences of pay-for-performance on the quality of care for minority patients. *Am Heart J*. 155(3):571-6.

La Caze, A. (2009). Evidence-Based Medicine Must Be... *Journal of Medicine and Philosophy*. 34:509-27.

Lim, W., Arnold, D.M., Bachanova, V., Haspel, R.L., Rosovsky, R.P., Shustov, A.R. and Crowther, M.A. (2008). Evidence-based guidelines--an introduction. *Hematology Am Soc Hematol Educ Program.* 26-30. doi: 10.1182/asheducation-2008.1.26.

Loughlin, M. (2008). Reason, reality and objectivity--shared dogmas and distortions in the way both 'scientistic' and 'postmodern' commentators frame the EBM debate. *J Eval Clin Pract.* 14(5):665-71.

Loughlin, M. (2009). The basis of medical knowledge: judgement, objectivity and the history of ideas. *J Eval Clin Pract*. 15(6):935-40.

McCartney, M., Treadwel, l J., Maskrey, N. and Lehman, R. (2016). Making evidence based medicine work for individual patients. *BMJ.* 353:i2452.

McKibbon, A, Hunt, D., Richardson, W.S., Hayward, R., Wilson, M., Jaeschke, R., Haynes, B., Wyer, P., Craig, J. and Guyatt, G. (2002). *Finding the evidence.* In *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice.* G. Guyatt and D. Rennie (eds). Chicago, American Medical Association.

Mercuri, M. and Gafni, A. (2011). Medical practice variations: what the literature tells us (or does not) about what are warranted and unwarranted variations. *J Eval Clin Pract.* 17(4):671-7.

Montgomery, K. (2006). *How Doctors Think: Clinical Judgment and the Practice of Medicine*. New York, Oxford University Press.

NICE. (2016a). *"Find guidance".* Retrieved 30 October, 2015, from https://www.nice.org.uk/guidance

NICE. (2016b). *"Standards and Indicators".* Retrieved 30 October, 2015, from www.nice.org.uk/standards-and-indicators

NICE. (2004). *Reviewing and grading the evidence. NICE: Guideline Development Methods (Vol. 7).* London, National Institute for Health and Care Excellence.

NICE. (2016). "*Depression in adults: recognition and management. Clinical guideline [CG90]".* Retrieved 29 April, 2016, from https://www.nice.org.uk/guidance/cg90.

Nordenstrom, J. (2007). *Evidence-based medicine in Sherlock Holmes' footsteps.* London, Blackwells.

Papineau, D. (1994). The virtues of randomization. *British Journal for the Philosophy of Science.* 45:437–50.

Petticrew, M. and Chalmers, I. (2011). Use of research evidence in practice. *Lancet.* 378(9804):1696.

Prasad, K. (2013). *Fundamentals of Evidence Based Medicine.* India, Springer.

Qaseem, A., Snow, V., Owens, D.K. and Shekelle, P.; Clinical Guidelines Committee of the American College of Physicians. (2010). The development of clinical practice guidelines and guidance statements of the American College of Physicians: summary of methods. *Ann Intern Med.* 153(3):194-9.

Randolph, A.G., Haynes, R.B., Wyatt, J.C., Cook, D.J. and Guyatt, G.H. (1999). Users' Guides to the Medical Literature: XVIII. How to use an article evaluating the clinical impact of a computer-based clinical decision support system. *JAMA.* 282(1):67-74.

Rapp, C.A., Etzel-Wise, D., Marty, D., Coffman, M., Carlson, L., Asher, D., Callaghan, J. and Whitley, R. (2008). Evidence-based practice implementation strategies: Results of a qualitative study. *Community Ment Health J*. 44(3):213-24.

Richardson, W.S., Wilson, M.C., Nishikawa, J. and Hayward, R.S. (1995). The well-built clinical question: a key to evidence-based decisions. *ACP J Club*. 123(3):A12-3.

Rosenberg, W. and Donald, A. (1995). Evidence based medicine: an approach to clinical problem-solving. *BMJ.* 310(6987):1122-6.

Sackett, D.L. (1986). Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest*. 89(2 Suppl):2S-3S.

Sackett, D.L. (1989). Inference and decision at the bedside. *J Clin Epidemiol.* 42(4):309-16.

Sackett, D.L. (1993). Rules of evidence and clinical recommendations for the management of patients. *Can J Cardiol*. 9(6):487-9.

Sackett, D.L., Haynes, R.B. and Tugwell, P. (1985). *Clinical epidemiology: a basic science for clinical medicine, First edition.* Boston, Little Brown.

Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B. and Richardson, W.S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ*. 312(7023):71-2.

Sackett, D.L. and Wennberg, J.E. (1997). Choosing the best research design for each question. *BMJ*. 315(7123):1636.

Sackett, D.L. (2000). The fall of "clinical research" and the rise of "clinical-practice research". *Clin Invest Med.* 23(6):379-81.

Shackelton RJ, Marceau LD, Link CL, McKinlay JB. (2009). The intended and unintended consequences of clinical guidelines. *J Eval Clin Pract.* 15(6):1035-42.

Schoenwald, S., Carter, R., Chapman, J. and Sheidow, A. (2008). Therapist adherence and organizational effects on change in youth behavior problems one year after multisystemic therapy. *Adm Policy Ment Health & Ment Health Serv Res*. 35(5):379-94.

Schumacher, J.E., Ritchey, F.J., Nelson, L.J. 3rd, Murray, S. and Martin, J. (1995). Malpractice litigation fear and risk management beliefs among teaching hospital physicians. *South Med J.* 88(12):1204-11.

Sekhar, M.S. and Vyas, N. (2013). Defensive Medicine: A Bane to Healthcare. *Ann Med Health Sci Res*. 3(2): 295-6.

Senn, S. and Harrell, F. (1997). On wisdom after the event. *J Clin Epidemiol*. 50(7):749-51.

Sleight, P. (2000). Debate: Subgroup analyses in clinical trials: fun to look at - but don't believe them! *Curr Control Trials Cardiovasc Med*. 1(1):25-7.

Stegenga, J. (2014). Down with the Hierarchies. *Topoi.* 33:313-22.

Straus, S.E, Richardson, W.S., Glasziou, P. and Haynes, R.B. (2005). *Evidence-based medicine: How to practice and teach EBM.* 3rd Ed. Edinburgh, Churchill Livingstone.

Straus, S.E., Tetroe, J. and Graham, I.D. (2009). *Knowledge Translation in Health Care: Moving from Evidence to Practice.* Oxford, Blackwell Publishing Ltd.

Summerskill, W. (2005). Evidence-based practice and the individual. *Lancet.* 365(9453):13-4.

Szatmari, P. (2004). Response to Dr. Gupta. *Evid Based Ment Health*. 7(4):97-8.

Tanenbaum, S.J. (2012). Improving the quality of medical care: the normativity of evidence-based performance standards. *Theor Med Bioeth.* 33(4):263-77.

The University of Arizona. [no date]. "*EBM searching*". Retrieved 13 April, 2016, from http://webtest.ahsl.arizona.edu/curriculum/pharmacy/Ebm_search.html

University of Illinois at Chicago. (2016). "*Evidence Based Medicine: PICO*". Retrieved 13 April, 2016, from researchguides.uic.edu

U.S. Preventive Services Task Force. (2008). *U.S. Preventive Services Procedure Manual*. AHRQ Publication No. 08-05118-EF.

Walker, H.K., Hall, W.D. and Hurst, J.W. eds. (1990). *Clinical Methods: The History, Physical, and Laboratory Examinations.* 3rd ed. Boston, Butterworths.

Web of Knowledge. (2016). *"Web of Knowledge".* Retrieved 04 December, 2015, from https://webofknowledge.com/

Wieringa, S. and Greenhalgh, T. (2015). 10 years of mindlines: a systematic review and commentary. *Implement Sci.* 10:45.

Woolf, S.H., Grol, R., Hutchinson, A., Eccles, M. and Grimshaw, J. (1999). Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. *BMJ.* 318(7182):527-30.

Chapter 3: Research improvements

3.1. Abstract

This chapter is focused on recent research improvements. My primary aim is to examine the prospects of three methodological refinements with respect to the problem of extra information (PEI). I shall argue that well-conducted *Pragmatic Randomized Controlled Trials*, *Subgroup Analyses*, and *n of 1 Randomized Controlled Trials* can do much to ameliorate standard problems that often jeopardize the adequacy of physicians' clinical recommendations. But even though these technical upgrades can aid physicians in obtaining probabilities for more fine-grained reference classes of patients, I shall contend that none of them eradicates the PEI and therefore do not eliminate the need for clinical judgment required by the Discretionary Approach (DA).

3.2. Recapitulation

In the first chapter I argued that, if the physician wants to maximize expected utility for her patient, the right probabilities are the probabilities in the reference class defined by everything she knows about the patient. The physician cannot avoid making a judgment about what these probabilities are; this is the role of clinical judgment and EBM is not doing the patient a favour by encouraging the physician to use probability estimates from valid research results and just assume that the extra features present in the patient will not have any probabilistic causal effect on the outcome of interest.

The second chapter offered a defence of my reconstruction of the EBM approach. As I explained, while it is true that the standard definition of EBM (Sackett et al. 1996) seems to leave room for the application of clinical judgment, I argued that the methods used to teach and implement EBM, along with various professional, economical and legal influences, reduce the latitude for the exercise of clinical discretion, and end promoting recommendations that neglect the PEI.

This chapter focuses on a line of argument that, if adopted by supporters of EBM, could provide them with a practical answer to the challenge of basing clinical decisions on the right probabilities for each individual without yielding ground to physicians' clinical judgment. This strategy does not deny my claims that the right probabilities for clinical decision-making are the probabilities based on everything the physician knows about the patient (*DA probabilities*) (§ 1.5), and that the EBM approach provides physicians with probabilities that ignore part of the available information *(EBM probabilities)*. However, this strategy aims to address these issues without appealing to clinical judgment. Instead it claims that suitable refinements in research methodology can in due course overcome the relevant evidential gap between *EBM probabilities* and *DA probabilities*, and therefore that EBM will eventually be able to offer patients recommendations based on the right probabilities, but this time not filtered or refined by physicians' judgment.

This chapter proceeds as follows. The next section provides the reader with the necessary background to understand in more detail what kind of obstacles hamper the generation of general efficacy claims (§ 3.3). This is important because I think that most refinements proposed by supporters of EBM, though addressing some of the challenges related to general efficacy claims and –albeit partially– helping EBM practitioners with the PEI, cannot be used to defend the thesis that clinical judgment is dispensable. Then, in the following section (§ 3.4) I discuss three research refinements: (§ 3.4.1) Pragmatic Randomised Controlled Trials (PRCTs), (§ 3.4.2) subgroup analyses (SGAs), and (§ 3.4.3) N of 1 Randomised Controlled

Trials (N1s). I shall contend that these improvements help physicians to address crucial challenges related to basic *statistical inference* (estimation), *causal inference* (confounding), *external validity* (extrapolation), and –to some extent– the PEI. Nonetheless, I shall insist that is unlikely that these methodological advances will be able to obliterate the gap between the probabilities provided by our best research methods (*EBM probabilities*) and the right probabilities for each individual (*DA probabilities*). For this reason, while they are steps in the right direction, they are not an effective means for dispensing with physicians' judgment altogether

Finally, the chapter concludes with a discussion of the prospects of *personalised* and *precision* medicine[44]. These types of medicine are certainly distinct from standard EBM practice (4S model and EBGs model) as discussed in (§ 2.4), but deserve consideration because they exploit the idea of using research refinements to offer patients individualized prescriptions, and their advocates often style themselves as practitioners working at the "cutting edge" of evidence-based practice (e.g. Kumar, 2011; Sharma, 2014). I shall maintain, however, that although these movements are right in thinking that current EBM recommendations need to be improved, they tend to overstate the level of individualization achievable by their methods and give the false impression that the adequacy of recommendations does not depend on careful clinical discretion. While I think that there is progress in tailoring prescriptions by using genetic makeup or sophisticated physiological variables, such kind of tailoring, far from eliminating the need for a considered judgment, makes its application even more relevant.

3.3. General efficacy claims

When introducing the rationale behind EBM recommendations in the context of therapy (§ 1.4.1), I assumed that conventional RCTs have the capacity to provide physicians with evidence for general efficacy claims such as *"topical antibiotics are not, on average, an effective treatment for patients with acute conjunctivitis"*.[45]

[44] Neither *personalised medicine* nor *precision medicine* should be confused with *person-centred medicine*. The main elements of the latter will be discussed in chapter 4 (§ 4.2.2).

[45] Let me stress that by "general efficacy claims" I mean both claims about the average efficacy or average "non-efficacy" of some intervention (pharmacological, psychological, etc.) for a population defined by membership to standard diagnostic reference classes (e.g. patients with major depressive disorder, patients with asthma, patients with diabetes type I, and so on). Notice also that that the equivalent to a general efficacy claim in the context of diagnosis and prognosis are general "diagnostic/prognostic claims" such as "shortness of breath is, on average, a sign of severe asthma" or "patients with acute conjunctivitis recover, on average, within one week".

At that stage I did not dwell on the inferential process leading to general efficacy claims, but simply limited myself to pointing out that, given certain assumptions, such kind of claims are in practice possible. However, in avoiding a detailed discussion about the rationale behind the kind of general efficacy claims supported by EBM, I did not want to convey the idea that the inferential process leading to such claims is without problems. [46] My reason for not dwelling on such problems was simply that I wanted to direct the reader's attention to the PEI, which is an additional challenge that physicians must face during the clinical encounter, and a challenge that may arise even if they have access to relevant RCTs, which provide them with sound general efficacy claims; that is, even if physicians were in a position where they had no doubt about the truthfulness of efficacy claim applicable to an intended target population their particular patients belong to.

As it happens, I am well aware about the difficulties of the scientific process leading to general efficacy claims. After all, it is precisely because determining whether, to what extent, and under what circumstances treatments are effective (or ineffective) for intended target populations is not an easy task that many generations of scientists, statisticians, methodologists, and philosophers of science, have worked and are continuing to work to improve our inferential methods, which includes the development of more rigorous research designs and more sophisticated approaches to data analysis (e.g. Matthews, 1995; Doll, 1992; Alyass et al. 2015; Porter, 1997; Marks, 1997, Feinstein, 1967, 1968; Auffray et al. 2016).

For the sake of clarity I shall distinguish between three inferential problems, which are connected to, but different from, what I call the PEI. This may be familiar ground for some readers, but rehearing these issues will facilitate exposition, and help me to explain why I think the set of research refinements put forward by supporters of EBM in the hope of bringing about truly individualised recommendations do not provide physicians with a definite solution to the PEI.

3.3.1. Basic statistical inference

The first problem that needs to be kept in mind is the perennial challenge of standard statistical inference. Under this term I include different types of estimation problems, which arise whenever one can obtain accurate estimates of the probability of a certain outcome of interest for a given target population using samples. The main concern here has to do with the influence of sampling fluctuations in the resulting estimates, which is, of course, a reasonable

[46] In "the risk GP model", Fuller and Flores (2015) consider the problems of general efficacy claims in the context of medical prediction in what we call the step of "generalization".

worry because, whatever probability a given estimate is for, it will not be a good estimate if it comes from an unlucky sample (Steyerberg, 2009; Hacking, 2001).

Note that the standard way to address estimation problems is with properly drawn large samples, so to decrease the chances that our sample is in one of the tails of the distribution by "bad" luck. This is the reason why both physicians and statisticians worry about issues such as sample size when estimating parameters and during hypothesis testing (e.g. Peto et al. 1995). For, other things been equal, it is generally reasonable to take more seriously estimates that come from large samples than those that come from small ones.[47] Note also –as this will be useful for the forthcoming discussion on subgroup analyses and N of 1 RCTs– that if we put estimation worries to one side, then we can say that the estimates obtained can be safely equated to the probabilities of interest for the population our sample is representative of. [48]

3.3.2. Basic causal inference

The second problem that affects general efficacy claims is the problem of causal inference. Without risking much simplification, a useful way to think about this problem, from the point of view of probabilistic causation and causal modelling, is as a challenge related to confounding factors (Hitchcock, 2010; Papineau, 2012). If everything has gone well with respect to basic statistical estimation, the aim of researchers is now to go from simple correlations to causal ones (that is, for example, from claims such as *"the probability of recovery is higher with topical antibiotics than without them in patients with acute conjunctivitis*" to claims such as *"topical antibiotics do not cause, on average, recovery in patients with acute conjunctivitis")*.[49]

It is worth noting, at least in passing, that causal inference is a topic of active debate among

[47] Of course, what counts as a "small" or "large" simple will depend on the hypothesis under investigation. In medical research, various elements are considered, including general factors such as the prior probability of the hypothesis, and more technical ones such as the type of estimator (e.g. mean, proportion, etc.).

[48] As a statistical aside, at certain point in the past medical researchers on the side of EBM seem to have failed to distinguish between the problem of estimation, which can be effectively dealt with by increasing the sample size, and the problem of external validity, which may remain even in studies conducted with very large samples. Since an accurate estimate is not the same as an applicable estimate for the intended target population, meta-analyses should not be considered as methodological refinements that necessarily help with the problem of external validity. In fact, as several authors have pointed out (e.g Borenstein et al., 2009) meta-analyses can provide physicians with estimates that are more accurate but less useful because it remains unclear what actual target population such accurate estimates are representative of. Many real meta-analyses are conducted with primary studies, which were in turn conducted with samples known to be unrepresentative of the intended target populations or with samples whose representative remains unclear because of bad reporting (e.g. Travers et al, 2007; Petersen et al, 2007). In consequences, the estimates obtained from such meta-analyses might be more precise, but in no way more informative because they inherit the problems of primary studies.

[49] As I pointed out in the first chapter, the question of whether a given correlation is truly causal or due to some confounding is not as important in the context of diagnosis and prognosis as it is in the context of therapy, for in the former contexts relevant non-causal correlations can guide clinical decision-making without problems.

medical researchers, statisticians, epidemiologists, and philosophers (e.g. Russo and Williamson, 2007, 2011;Vandenbroucke et al. 2016; Hernán and Robins 2006; Beebee et al. 2010). In this regard, the EBM movement has been widely criticised for adopting an overenthusiastic attitude towards RCTs (e.g. Worrall, 2007, 2011; Cartwright, 2007). From my perspective, overenthusiasm towards RCTs is not questionable because it rooted in alleged virtues that are not real (e.g. the capacity of randomization to provide a probabilistic guard against unknown confounders). Overenthusiasm towards RCTs is pernicious because it neglects the weaknesses of RCTs and fails to recognize that causal inferences can also be reached by alternative methods that may offer additional advantages over RCTs and do not deserve to be categorically classified as "second class" citizens. Regrettably, this has been the case with methods such as observational studies and basic research on mechanisms, which have strengths that conventional RCTs do not possess (e.g. Worrall, 2007, 2011; Cartwright, 2007; Clarke et al, 2014) and which have not been sufficiently reflected in EBM's standard rules of evidence (§ 2.4.1.3 and 2.4.1.4). Nonetheless, we need not discuss the many facets of this debate here. For my purposes it is only necessary to stress that I accept that we can learn about causal relationships, in a technical sense, via different routes including both systematic research (e.g. RCTs, observational studies, basic scientific experiments) and non-research methods (e.g. clinical experience and causal reasoning based on mechanisms) each of which has its own strengths and weaknesses. In this regard, it is worthwhile stressing that while conventional RCTs can ensure that the intervention under investigation "causes" the outcome of interest in the trial population, they cannot, without a number of additional assumptions, ensure that the intervention will, on average, cause the outcome if applied to the target population.[50] As Nancy Cartwright (2011), puts it, an *"ideal RCT"* can tell us that a certain treatment *"works somewhere"* but it will not, by itself, tell us whether it will work if applied outside the trial population, for instance, if applied to the intended target population in standard clinical settings (p.1401).

Furthermore, with regard to the interpretation of general causal claims in the context of medicine focused on individuals, it is worth emphasising that I am happy to accept, along with supporters of EBM, that RCTs, if well executed, can be excellent methods to discover causal relations that are applicable to particular patients. However, I neither think that RCTs are the only method to discover causes for individuals, nor that when general causal relations

[50] The technical sense of probabilistic causation I have in mind here assumes that an intervention I is a cause of the outcome O if and only if there is some context in which I fixes and above-average single-case objective probability for the outcome O. This is equivalent to say that intervention I causes O in the trial population just in case there are at least some types of patients for which the Prob (O|I) > Prob (O|¬I) (See, Papineau 1985, 1989, 1994). Note that this account does not take *causal unanimity* as a requirement to establish causation, that is, I do not assume that in order to conclude that I causes O, Prob (O|T) > Prob (O|¬T) must be true in all contexts. (The details of this stronger notion of causation can be found in Eells, 1987 and Cartwright, 1989.)

have been established via RCTs, such causes necessarily have clinical applicability to individuals. What I mean to say here is that even when a physician is certain that a general causal claim holds true for a certain standard target population, and she knows that her patient belongs to this target population, she may still think that such general causal claim ought not to be transformed into a *particular efficacy claim* for the patient in question[51]. As I argued in the first chapter (§ 1.4.2), extra information in the form of interaction effects can either reverse or amplify causal effects for individuals; and before physicians transport efficacy claims from research to individuals they will do well in exercising their judgment to estimate the potential effects of such information on the outcome of interest (§ 1.5).

3.3.3. External validity

At the end of the first chapter (§ 1.6.2.1) I offered a few remarks on the problem of external validity, which I contended is a serious issue for conventional RCTs and has been underestimated by some supporters of EBM [52]. For purposes of this chapter I shall limit myself to stress that even if basic statistical inference and causal inference were not a problem, objections to the external validity of research findings can downgrade or even invalidate the practical import of general efficacy claims, even if they are otherwise methodologically flawless (See Rothwell, 2006; Steckler and McLeroy, 2008; Pearce et al, 2015; Cartwright, 2013).

The inferential difficulty underlying the problem of external validity has been identified using various terms which, regrettably have not been defined as precisely as one would wish. For example, authors concerned with this topic in the context of EBM often use terms such as the problem of extrapolation, the problem of transportability, the problem of lack of representativeness, or sufficient similarity assumptions, each which emphasise aspects of this problem from slightly different perspectives (Khorsan and Crawford, 2014).

Nonetheless, issues of terminology aside, one can make sense of the essential difficulty behind the problem of external validity as follows. Consider a well-conducted RCT: Having followed estimation procedures according to standard rules, as well as having randomized treatment allocations properly, now one can say that researchers have good reason to believe that, say, the correlation between the treatment and the outcome of interest they have found implies that there is some context in which the treatment makes a causal difference to the outcome of interest (Papineau, 1994). However, even if researchers are quite certain that the

[51] As Fuller and Flores (2016) argue in his Risk GP Model, *particular efficacy claims* require a step of "particularisation", which is based on assumptions that might not be easily met.

[52] For example, Jeremy Howick (2011), citing research conducted by Vist et al., (2008) –which in truth is of questionable relevance to the problem of external validity (See Fuller, 2016)– asserts that *"even if randomized trials appear to involve unrepresentative populations, the results apply to the target population"* (p.43).

have obtained a precise estimate which can given a causal interpretation of the kind "treatment T causes, on average, outcome O", they might still remain unclear about what the actual target population in which T makes this causal difference to O.

Of course, the obvious question is why researchers are unclear about the actual target population their general efficacy claim is representative of. A simple answer, which does not sacrifice much precision, is that the samples with which conventional RCTs are conducted are almost invariably *not* random samples of the intended target populations. [53] So, for example, even if we have an RCT conducted in a sample of subjects defined by the presence of a diagnosis of type II Diabetes, if the sample was not properly drawn from the intended target population[54], we will not know which among the *actual populations* which have been defined by the diagnosis of type II Diabetes this sample is representative of.

Note also that, from the clinical point of view, the problem of external validity can be understood as a general challenge for the generalization or applicability of research findings, for it raises doubts about the extent to which general efficacy claims can truly support recommendations for standard target populations in actual clinical settings (Rothwell, 2006, 2010; Pearce et al. 2015). [55] Thus, since external validity is a serious issue for the prospects of the EBM movement, understood as an extensive programme to improve the quality of medical care, methodological improvements to obtain research findings with increased levels of transportability can be found among the standard solutions to address this challenge (Treeweek and Zwarenstein, 2009; Patsopoulos, 2011).

So, to sum up, I have mentioned three problems that are directly related to the question of what general efficacy claims research findings can tell us about, namely, standard statistical estimation, causal inference, and external validity. These problems are relevant because the

[53] For the problem of external validity be understood as a problem of representativeness, the term sample should be read as encompassing both the members of the sample and also several characteristics of the experimental set-up such as setting, treatment regimens, follow-up periods, etc. We will return to this point later on in this chapter when discussing pragmatic RCTs (§ 3.4.1).

[54] The frequentist and bayesian schools of statistical inference have different views about the virtues of random sampling. Authors typically identified with the Bayesian School tend to accept that random sampling carries certain technical advantages, they do not think it necessary for sound inferences (See Howson and Urbach (1989); Berry (2006). The position of frequentist statisticians is, at least in some sense, odd, for while they normally defend the importance of random sampling, when comes to particular studies where samples had not been randomly obtained sometimes they ignore this and interpreted parametric statistical tests as if had been obtained from random samples. For the frequentist position on random sampling see Mayo and Spanos (2010).

[55] Note that there is a sense in which the problem of external validity is much more than an academic issue concerning statisticians and medical methodologists. Several authors have pointed out that with the advent of EBM movement pharmaceutical companies have misused RCTs as marketing tools to promote recommendations based on efficacy findings that were obtained in samples conveniently arranged with patients of good prognosis, which are very different from those conforming intended target populations in real settings (e.g. Ioannidis, 2016; Greenhalgh et al., 2014; Christmas, 2014; Rothwell, 2006).

truthfulness of general efficacy claims depends on how well the aforementioned difficulties have been dealt with [56], and more importantly, because if they are not addressed adequately the resulting efficacy claims might lead to incorrect clinical recommendations. But, as the reader now knows, even if general efficacy claims are sound, there is an additional problem for the physician, a problem that we have not yet started to worry about in this chapter, but a problem which has to be faced if the physician wants to maximize expected utility for each particular patient: the PEI. To put it simply, even when the three previous problems have been satisfactorily dealt with, and the EBM practitioner is happy to accept that the research findings she knows about can back up general efficacy claims, the physicians ought not to ignore the extra information she has about the patient and ought to exercise her judgment to estimate the right probabilities. As I discussed in the first chapter (§ 1.5), my position regarding the PEI is that it is unreasonable to ignore it, for physicians who assume that their extra information is probabilistically irrelevant will fail to estimate as best as they can the probabilities that maximize expected utility for their patients.

However, in this chapter we will consider an alternative attitude towards the PEI. The thought is the following: If previous methodological developments have helped researchers to address the challenges of statistical and causal inference, why one would not think that the same or new methodological improvements can free physicians from the PEI? The logic it is as simple as it is compelling: if throughout the years the collective efforts of methodologists and statisticians have led to (a) improvements in statistical estimation, (b) a better capacity to deal with the problem of confounding causes and (c) arguably more transportable research findings, then it does not seem unreasonable for supporters of the EBM movement to believe that research improvements could do the same with (d) the PEI. More precisely, could not the introduction of new research designs such as *pragmatic controlled trials*, more sophisticated *sub-group analyses*, and better and more numerous N = 1 RCTs bridge the evidential gap between research and individuals? Perhaps this kind of improvement can, as the supporters of precision medicine have sometimes promised: *"predict the treatment effect for individual patients based on all relevant characteristics together"* (Dorresteijn et al. 2011 p.2). And if these authors are right, then EBM could fulfil its aim of basing recommendations on objective probabilities estimated via the best research evidence without yielding ground to clinical judgment.

[56] Recall that in the context of therapy efficacy claims require sound statistical and causal inference as well as a good level of external validity, but since the claims necessary to support recommendations in the context of diagnosis and prognosis do not require sound causal inference, researchers typically focus on good estimation (e.g. Hemingway et al., 2013; Riley et al., 2013) and –more recently–on ensuring better external validity (e.g. Bleeker et al., 2003; Collins et al., 2014).

Below I shall consider the aforementioned research improvements, which I admit should be regarded as steps in the right direction. However, I shall argue that, in the end, none of them makes the exercise of careful clinical judgment superfluous. Although PRCTs, SGAs and N = 1 RCTs will typically deliver probabilities closer to the ones physicians are looking for, it is unlikely that these methodological advances will be able to collapse the difference between the probabilities provided by the best research findings and the probabilities needed for individual decision-making.

3.4. Research improvements

3.4.1. Pragmatic randomized controlled trials

The randomized controlled trial (RCT) is a notable research design. RCTs are often ascribed the capacity to provide physicians with research findings that are "internally valid", a term used to convey –in this particular case– a combination of robust statistical estimation and control for confounding factors, including unknown ones.[57] But RCTs, even those perfectly conducted, have important weaknesses. Some of these limitations are not necessarily inherent to this research design, but, since they are present in most instances, they acquire high practical significance. Of the three problems mentioned in the previous section, conventional RCTs are particularly vulnerable to the problem of external validity. As I said before, this problem is different from the problem of extra information (PEI), and a problem that has captured philosophical interest on its own (Cartwright, 2009, 2010; Rothwell, 2010; Howick, 2011; Howick et al. 2013).

One reason to be interested in this problem is that the level of external validity of the evidence endorsed by EBM is important to assess the overall cogency and practical prospects of this movement. For if EBM is understood as a general programme to modify medical care on the basis of our best research evidence, then the success of this approach certainly depends on the assumption that what EBM regards as the best research evidence (RCTs or meta-analyses of RCTs for questions of therapy) is typically –and perhaps also sufficiently– externally valid to be applicable to standard clinical populations. Nonetheless, this is not the reason why I am interested in the problem of external validity here. I am mentioning this problem here because research improvements aimed at addressing the problem of external validity could easily be misinterpreted as solutions to the problem of individualisation. And it

[57] This, of course, need not be the case, RCTs can be conducted in very small samples and as a result provide weak evidence for correlations. Furthermore, if randomization is not properly carried out, RCTs do not offer any warrant to interpret probabilistic differences as causal differences. Nonetheless, it is commonly assumed that researchers in charge of conducing conventional RCTs try to ensure that both statistical estimation and causal inference be as robust as possible.

would be unfortunate if valuable strategies to deal with the lack of external validity of RCTs end up being discredited because they do not solve the PEI or, which is perhaps even worse, end up being undeservedly regarded as solutions to the PEI, which they are not.

The Pragmatic Randomized Controlled Trial (PRCT) is among the research improvements, whose advantages might be wrongly interpreted as a solution to the PEI. The PRCTs are a sensible refinement over conventional RCTs, which as I said are normally affected by the problem of external validity. As several authors have pointed out (See e.g. Godwin et al. 2003; Fortin et al. 2006; Uijen et al. 2007; Travers et al. 2007; and Rothwell 2005a, 2006), the practical import of RCTs can be very limited because these trials usually investigate the efficacy of treatments under experimental circumstances very difficult to replicate in real settings. This implies that the findings from conventional RCTs can be misleading because they were obtained with samples that are not representative of the target populations. And if so, although researchers can claim that there is some context where the treatment under investigation is effective for some patients, researchers left physicians without knowing what the actual target population the findings are supposed to be representative of.

It was against this backdrop that PRCTs emerged as a feasible alternative to increase the level of external validity of the findings of conventional RCTs (e.g. Johnson et al. 2014). [58] For, in essence, what was needed was to carry out RCTs with samples more akin to standard clinical populations, and to modify RCT protocols so that treatment administration regimes and follow-up periods resemble those that were practically possible in actual settings (Hotopf, 2002; Macpherson, 2004).

The PRCT has proven to be a useful research improvement. Characteristics such as less stringent inclusion and exclusion criteria, feasible treatment protocols, and the use of clinically relevant measures (e.g. measuring outcomes of interest for real patients within an adequate time framework) normally result in estimates of higher external validity. So, the main advantage of PRCTs is that the actual target population the findings are representative of is easier to identify, and as a result, the findings can be more confidently extrapolated to the corresponding target population in real settings (Marks et al. 2009).

Now, if we consider an efficacy claim, say "antidepressants are effective to treat major depressive disorder" and we are told that this claim is backed up by a PRCT rather than a conventional RCT, it becomes apparent that, other things being equal, a general efficacy

[58] I use the term "feasible" because most statisticians would accept that random sampling would be a useful method to ensure the representativeness of the sample in a statistical sense--but the practical difficulties of carrying out such kind of sampling procedures are often daunting.

claim supported by a PRCT has better prospects in terms of applicability than that supported by a conventional RCT, for such claim is more likely to hold true in the intended target population when backed up by the former than by the latter.

However, the benefits of the PRCT with respect to the problem of external validity should not be confused with a potential advantage with respect to the PEI. The PEI, unlike the problem of external validity, is a problem that emerges when the physician has additional information about a particular individual member of the target population. (Recall that in the therapeutic clinical case presented in the first chapter Mr Smith was a particular member of the general reference class "patients with acute conjunctivitis".) Thus, even if the physician had access to a perfectly conducted PRCT; that is, even external validity were out of question, the presence of extra facts about the particular member of the target population the physician is interested in remains a crucial challenge that, as I argued, cannot be ignored if the physician wants to arrive at the right recommendation. So, the point is that, from the perspective of the EBM approach, which encourages EBM practitioners to issue recommendations based on the best research available, the fact that such research is now instantiated by a PRCT rather than a conventional RCT does not change the fact that this approach still directs the physician's attention to the wrong probabilities and neglects the PEI.

So, although an ideal PRCT could successfully bridge the so-called gap between *research* and *practice* understood as a gap between research samples and actual target population (Chalkidou et al. 2012; Longford and Nelder, 1999), this research refinement does not bridge this gap if it is understood as a discrepancy between the probabilities for which there is valid research available and the right probabilities for clinical recommendations. For this reason, the PRCT cannot be considered a solution to the PEI.

Thus, if I am right in that the correct probabilities for clinical practice are the probabilities based on everything the physician knows about the patient, the ideal PRCT will give a physician strong reasons to accept that there is certain therapy that, on average, is effective for members of a target population to which her patient belongs. But if the physician knows more about her patient, then the PRCT will not necessarily provide him with the probabilities she needs to decide whether to recommend or withhold such therapy in patient's best interest. Of course, as with conventional RCTs, PRCTs may sometimes deliver the right probabilities for clinical decisions, but this is something that depends on how much knowledge the physician has about her patient and is in no way related to the methodological changes that transform a conventional RCT into a PRCT. Careful exercise of clinical judgment in the light of everything the physician's knows about her patient remains, therefore, an unavoidable task for the rational physician, even in the presence of sound estimates from PRCTs.

3.4.2. Subgroup analyses

In the medical literature Subgroup analyses (SGAs) are commonly presented as strategies to explore the heterogeneity of outcomes within study samples (Sculpher, 2010; Keene and Garret 2014), and less often, as methods to avoid errors because of the low relevance of aggregate estimates to some members of the target population (Sun et al, 2014; Oxman and Guyatt, 1992; Rothwell, 2005b).[59]

Unlike PRCTs, which help physicians with the problem of external validity but do not address the PEI, SGAs are research refinements that permit recommendations based on probabilities for more fine-grained reference classes, and in virtue of this advantage, they may help EBM practitioners with the problem of extra information (PEI).

Nonetheless, although I think that SGAs are particularly valuable tools, which increase the relevance of research findings to individuals, I shall argue that they should not be interpreted as providing a definite solution to the PEI. Even in cases where the best research evidence available comes from SGAs, it would be wrong to think that SGAs eliminate the PEI. EBM practitioners should not take the probabilities from SGAs as equivalent to the right probabilities for individuals. As with the probabilities obtained from conventional RCTs, the probabilities obtained via SGAs may but need not be the right probabilities for particular patients. So, while I acknowledge that SGAs constitute a step in the right direction, advocates of EBM should not use them to argue that this approach can deal with the PEI without resorting to clinical judgment.

Before proceeding, a caveat is in order. SGAs are methods that have generated a great deal of controversy among statisticians, epidemiologists, and medical researchers (See e.g.: Brookes et al. 2001, 2004; Wang et al. 2007; Pocock et al. 2002; Jones et al., 2011; Naggara et al., 2011). It is not surprising, then, that one can find supporters and detractors of SGAs within the EBM camp (Sun et al. 2012, 2014; Gabler et al., 2009; Senn and Harrell, 1997; Altman, 1998). However, while it is worth noting that not every advocate of EBM would accept evidence from SGAs as the best evidence available, I can by-pass the technical details of this debate because they are largely orthogonal to the point I shall make in this section. The

[59] This section is entitled "subgroup analyses" but it should be read as encompassing several analytical tools normally classified under the heading of *"stratified medicine"*. (See Feinstein, 1998; Brookes et al., 2001, 2004; Shaw and Johnson, 2012). Accordingly, my use of the term subgroup analyses includes both (a) analyses that account for one additional variable at a time and also (b) more sophisticated techniques, which rely on various modelling strategies to account for the heterogeneity of treatments effects (Kent and Hayward, 2007ab, Kent et al., 2010; Wang et al., 2007; Peto, 2011; Hayward et al., 2006; Dorresteijn et al., 2011). It is worth noting, however, that the main difference between conventional SGA and more elaborated techniques such as test of interaction, which are favoured by supporters of EBM (Kent et al., 2010; Brookes et al., 2001, 2004) has to do with issues of estimation, which are important, but ultimately orthogonal to my point. The precise approach to explore interaction effects is not important to my point as long as the result is to obtain a probability for more refined reference classes.

reason for this is that the controversy about SGAs has to do with basic estimation issues that arise in the context of statistical inference (§ 3.2.1). Since estimation problems affect any approach to inference, regardless of the presence of the PEI, I shall put them to one side so to assess the ideal prospects of this research improvement. Of course, I am perfectly aware that the extent to which estimation problems hamper inferences based on SGAs is of practical importance for clinical medicine. But, again, since my argument contends that SGAs would not offer a definite solution to PEI even if estimation problems did not exist, we can facilitate exposition by considering the prospects of SGAs with respect to this problem under the admittedly unrealistic assumption that SGAs provide physicians with probabilities for actual target populations.

3.4.2.1. Fine-grained probabilities via SGAs

It is high time for me to address the question of to what extent SGAs address the PEI. Let us then, recall once more the case of Mr Smith from the first chapter, the patient diagnosed with acute conjunctivitis (AC), who was recommended not to use topical antibiotics since current EBM guidelines indicate that they are not effective (Steeples and Mercieca, 2012) (§ 1.4.2). Suppose, as we did before, that the treating physician, Dr Jones, learned that Mr Smith has AC and a history of Repeated Sexually Transmitted Diseases (RSTD). As I explained when I introduced this case –and expanded later when discussing the 4S model to teach EBM (§ 2.4.1.4)–, since the relevance of the presence of RSTD to the outcome of interest is difficult to defend according to EBM rules, it is reasonable to think that an EBM practitioner like Dr Jones would prefer to neglect the PEI, and assume that Mr Smith's history of RSTD does not make any difference to the outcome of interest.

By now, the reader will know very well that, even if we assume that estimation causal inference are not a problem, the evidence available can back up a general efficacy claim, which is not necessarily the right recommendation for Mr Smith. After all, if Dr Jones knows that Mr Smith has AC and RSTD, she cannot conclude that what on average is true for subjects with AC will also be true for patients with AC and RSTD. These are probabilities for different reference classes, and their difference is important because RSTD might interact with topical antibiotics so to turn them more effective, less effective or might not interact at all.

This simplified case illustrates that a SGA for subjects with AC and RSTD would be a very good thing to have for both Dr Jones and Mr Smith. For, at least in this simplified clinical scenario, such SGA would provide the physician with the right probabilities for clinical inference, or more precisely, the probabilities for a patient just like she knows Mr Smith to be.

Thus, since SGAs focus on specific subsets of patients, it is natural to think of this research improvement as a valuable tool, which provides physicians with more refined outcomes than main analysis from conventional RCTs. In doing so, SGAs effectively decrease the gap between general efficacy claims backed up by RCTs, which typically tell physicians what works, on average, for general classes of patients, and individual efficacy claims, which tell the physician what works, on average, for members of the reference class conformed by everything the physician knows about a particular patient (§ 4.1).

3.4.2.2. SGAs and the PEI

At this stage some readers might feel enthusiastic about the prospects of SGAs with respect to the PEI. Is it not plausible to think that SGAs, in due course, will collapse the difference between the probabilities for which valid research data is available (EBM probabilities) and the probabilities based on everything relevant the physician knows about the patient (DA probabilities), and when that time comes, clinical medicine might be able to dispose of clinical judgment and its accompanying perils?

However, even if by now it has become apparent that SGAs may offer more refined probabilities than conventional RCTs, it is not immediately obvious that SGAs will provide EBM practitioners with a definite solution to PEI. The main concern here is with the practical capabilities of both SGAs and individual physicians to gather relevant information. For us to take this possibility seriously one need to think about whether –and if so, how often–SGAs would be truly able to deliver probabilities that are individualised enough to match the right probabilities?

Since the answer to these questions depends on the capacity of SGAs to account for further information as well as on the amount of information physicians normally obtain from their patients during the clinical encounter, it will be useful to try to picture a clinical case in the most realistic way possible.

Think once more about Mr Smith's case and try to form a mental image of a real clinical situation. Without idealizing Dr Jones's abilities to conduct clinical interviews, it is plausible to assume that within the short time frame allowed for clinical encounters, she will be able to collect information about Mr Smith which might include (i) Mr Smith's personal history of efficacy of antibiotics under similar circumstances, (ii) previous adverse reactions or complications without therapy, (iii) vital signs, and (iv) other observations from the eye

examination such as redness grade and type of discharge [60]. As I argued in the first chapter, the right probabilities for Mr Smith are relative to this informational set, not relative to the part of this set for which valid clinical trials are available. And although it is correct that the probability for a patient with AC and RSTD is a probability that is closer to the probability of interest, given the amount of information normally available to physicians it is not a probability that is close enough to the right probability for the patient[61].

Of course, the enduring presence of a gap between these probabilities does not imply that SGAs are not useful. As I said, normally SGAs permit more fine-grained recommendations and will decrease the number of idiosyncrasies that need to be taken into account to estimate the right probabilities. But the point is that even in the presence of probabilities from SGAs it is likely that the physician will have to exercise her judgment to determine whether factors such as "redness grade", "type of discharge", "particular vital signs" or "previous response to similar treatments" modify the probability of the outcome of interest.

Finally, note that it remains possible that after considered judgment in the light of what the physician knows about the patient, the treating physician concludes that the extra features present in the patient make no probabilistic difference to the outcome of interest. But let me emphasise that since the PEI is a procedural worry rather than a consequential worry, what matters is not that the exercise of clinical judgment results in revised probability estimates for the patient, but rather that the physician does not put her judgment to one side and assume that the extra features present in the patient are negligible just because there is no valid research data supporting their relevance to the outcome of interest.

So, even if SGAs provide the EBM practitioner with more refined probabilities, which might account for a greater amount of the information gathered by the physician during the clinical encounter, the limitations that characterise estimates from conventional RCTs remain present, in SGAs. This implies that the gap generated by the presence of further information about the patient cannot be ignored and suggests that SGAs offer a valuable but partial solution to the PEI. For this reason, EBM practitioners should bear in mind that research findings from SGAs should not be automatically extrapolated to individuals, and such findings do not free the rational physician from her responsibility to estimate, as best as possible, the right

[60] Standard medical textbooks, such as the Oxford Handbook of Clinical Specialties (Collier et al., 2013), classify part of the information I have mentioned as standard elements of the clinical assessment. Furthermore, being a practicing physician myself I know very well that even the busiest (or "laziest") of my colleagues would ask their patients a minimal set of questions, which include personal information, history of present illness, previous use of medications and complications, and findings from physical examination.

[61] It must be noticed that at least some authors in the EBM camp have recognised that SGAs at typically based on *"univariable analyses [which] do not fully incorporate all available patient characteristics…"* (Dorresteijn et al., 2011, p.2) See also, Rothwell, 1995 and Kent and Hayward, 2007ab.

probabilities for her patient.

3.4.3. N of 1 RCTs

In this section we shall focus on Randomized Controlled Trials conducted in individual patients (N1s). My aim is to analyse the merits of N1s with respect to the problem of extra information (PEI). I shall argue that the N1 is a useful research design, which may lead to improved prescriptions for certain individuals, but does not provide a definite solution to the PEI. Even when the best research evidence is instantiated by findings from N1s, clinical judgment will normally be necessary to estimate the right probabilities for particular patients and therefore remains crucial to ensure the adequacy of clinical recommendations.

As their name indicates N1s share properties with two classes of research designs: RCTs and single case studies. The history of both RCTs and single case studies precedes the emergence of the EBM movement (See Hacking, 1988; Hill, 1952; Barlow and Hersen, 1984; Bothwell and Podolsky, 2016; Gabler et al., 2011; Fisher, 1990). But it would not be far from the truth to say that N1s, in their standard form, are a product of this movement (Guyatt et al. 1986).

Supporters of EBM have described N1s as the ultimate research design to investigate therapeutic efficacy in particular individuals (Guyatt et al. 1988, 1990). In contrast with conventional RCTs, whose focus is the generation of general efficacy claims for standard populations, N1s are typically (but not exclusively[62]) expected to produce particular efficacy claims; that is, claims about the efficacy of interventions in individual patients (Kravitz et al., 2008).

N1s are held in high regard among advocates of EBM. For example, Kravitz and colleagues (2008) stress the virtues of N1s by telling us that: *"…randomized, blinded single-patient (n-of-1) trials are uniquely capable of establishing the best treatment in an individual patient."* (p.533). And it is perhaps because supporters of EBM believe that N1s fulfil this role almost to perfection that in some hierarchies of evidence N1s have been ranked at the top, even above *"systematic reviews of randomized trials"* (Guyatt et al., 2000 p. 1292).

Since the reader might not be familiar with N1s, before explaining why I think this research design does not solve the PEI, a brief description of their methodology will be instructive. N1s are typically applied to investigate the efficacy of medications in the context of chronic diseases (Duan et al. 2013). The management of chronic diseases has two characteristics that are important to understand the role assigned to N1s. The first one is that poly-pharmacy, the use of several medications simultaneously, is the rule rather than the exception (Mannuci et

[62] For example, see Senior et al., 2013.

al., 2014); and the second one is that the standard therapeutic aim in this context is not full recovery but rather symptom reduction, or increments in quality of life (Parek et al., 2011). Thus, N1s are normally used to decide whether adding or eliminating a certain medication from a given therapeutic scheme is beneficial or harmful for a particular patient (Kravitz et al., 2014).

In N1s the patient receives both the treatment under investigation and the control intervention –which could be another active treatment or no active treatment. This is possible because N1s are crossover designs, in which interventions are administered sequentially, often in pairs separated by "wash out" periods (e.g. A-B, B-A, B-B, and so on) (Vohra et al, 2015).

N1s are considered among randomized research designs because the sequence of treatment allocation is established using a random device and concealed from both the physician and the patient. Normally, N1s require a research assistant (usually a pharmacist) who is responsible for preparing similar interventions so neither the patient nor the physician can evaluate the symptoms without being able to guess the treatment sequence. When the trial ends, which is typically after 3 pairs of treatment episodes but varies from one set-up to another, an estimate of the effect of each intervention with respect to the outcome of interest is calculated. Finally, a statistical test is often used to compare the estimates obtained between each other. The resulting findings are taken to be evidence about the efficacy (or non-efficacy) of the main treatment with respect to the control intervention (Guyatt et al., 1990; Kravitz et al., 2014).

After this brief description the reader might have several concerns about N1s, which probably range from the ethical to the methodological. For example, are N1s ethically justified in clinical contexts? And, under what circumstance is an attempt to establish the causal effect of a specific part of a composite treatment worth risking a decrease in the patient's health? [63] Or perhaps, with respect to the challenge of estimation (§ 3.3.1): how reliable are the estimates that can be obtained with N1s? Are not N1s typically based on too few observations? Or maybe, with regard to causal inference (§ 3.3.2), is randomization, as carried out in N1s, good enough guard against confounding?

All these questions are of practical clinical importance. But to address them would take us too far from the scope of this chapter. More importantly, since my main claim is that N1s can be

[63] Ethical discussions about N1s usually come to a point where it is necessary to ask explicitly what is the intention of the experiment. While an N1 might be ethically justified in the expectation that it might permit eliminating a medication for the patient which does not have an active role in the reduction of symptoms but which carries various side effects, the same N1 might not be ethically justified if the aim is to clarify whether the medication has an active causal effect for purely academic or scientific purposes, which are disconnected from the patient's benefit. See Kravitz et al., 2014 for a discussion focused on N1s, and Lilford and Jackson, 1995 for a general comment on the ethics of RCTs.

useful but do not solve the problem of extra information (PEI), N1s' estimation problems and their limitations with respect to causal inference are largely orthogonal and we need not dwell on them. Even in the ideal scenario where N1s provide the physician with true causal correlations, I shall still contend that this kind of research improvement does not make the exercise of clinical judgment redundant.

The best way to illustrate my point is with an example, for which I would like to consider a case adapted from the article where N1s were first proposed (Guyatt et al. 1986). John is a 13 years old patient who has suffered from chronic asthma for 6 years. When uncontrolled, John has several symptoms including chest tightness and shortness of breath during normal activities, as well as attacks where the symptoms get worse, even when she is at rest. Over several months, her physician adjusted the treatment regimen until John reported being significantly less symptomatic while taking a bronchodilator (Albuterol) and an anticholinergic agent (Ipratropium). Since the addition of Ipratropium was correlated with the stabilization of John's symptoms, her physician suspected that this medication might be causally related to his improvement. However, neither John nor her physician were convinced about this, for they were aware that Ipratropium could be correlated with another unknown factor, which could be causing the improvement in his symptoms.

Determining whether Ipratropium was causally responsible for his improvement was important for John for several reasons, including economic costs and, also because this medication carries side effects such as bladder pain, constipation and blurred vision (Ziebach et al., 2001). Thus, the physician and John agreed to carry out an N1 to find out whether Ipratropium was causally effective and thereby be able to tailor John's therapeutic scheme to better meet his needs.

Assume that the trial was conducted without technical problems, that is, John took the medication and the physician monitored his clinical status throughout the trial as expected. As **table 1** (see below) shows, the results of John's N1 trial suggest that there is a correlation between Ipratropium and a decrease in Johns' symptoms, and since this was a well-conducted randomized trial, this results are suggestive that Ipratropium, on average, causes symptomatic improvement for John. Now, if we put potential concerns with statistical estimation and potential problems with causal inference aside, we can take it that this N1 trial indicates that Ipratropium is, on average, effective for John. And so, according to what EBM regards as gold standard evidence for particular efficacy claims, the right recommendation for John ought to be to continue Ipratropium. But now, the central question of this dissertation arises once more: did this N1 truly solve the PEI for John's physician? Or, in other words, is this EBM individualized prescription for John based on the right probabilities?

Table 1: John's N of 1 RCT †

		Period 1	Period 2	Period 3	Period 4	Mean score
Pair 1	**Control**	4.43	4.00	4.14	4.29	**4.22**
	Ipratropium	4.43	4.86	4.71	4.71	**4.68**
Pair 2	**Control**	3.86	4.00	4.29	4.14	**4.07**
	Ipratropium	4.57	4.89	5.29	5.29	**5.01**
Pair 3	**Control**	3.71	4.14	4.43	4.43	**4.18**
	Ipratropium	4.29	5.00	5.43	5.43	**5.04**

Benefit is measured with a shortness of breath scale (Higher scores represent less symptoms). A paired t-test applied to the mean score pairs (4.68 and 4.22, 501 and 4.07, and 5.04 and 4.18), using two degrees of freedom, results in a t value of 5.07, which has a corresponding **p value** of **0.037.** † Adapted from Guyatt et al. 1986 and 1988.

Some readers (and, I suspect, many physicians) might be a bit puzzled at this point. Is it not obvious enough that there is no further individualization possible for John than the one already accomplished by the above-described N1 trial? If the N1 was conducted in John himself, how is it possible that the physician could know something extra about John that is not properly accounted for in the N1?

The crucial point to stress here is that PEI is not a problem about finding the most individualised probabilities valid research can offer, but rather the problem of obtaining the probabilities fixed by everything the physician knows about her patient.

With this distinction in mind, it can be appreciated that the N1 successfully solves the problem of providing the best possible estimate research can provide for an average version of John at the time of decision-making. But since the PEI has to do with using all information available, regardless of its source, it is quite clear that during or after the N1 the physician might still learn something extra about John, something that may change the probability that Ipratropium causes a reduction of symptoms. For example, the physician might learn that John now has a viral infection, whose pharmacological treatment interacts with Ipratropium so to decrease its effect. Or perhaps that John has made a recent dietary change, which soon will affect the absorption of Ipratropium and therefore its effect on the symptoms.

So, although supporters of EBM are right that it is reasonable to assume that N1 decreases the number of probabilistically relevant additional factors that the physician might know about her patient, the fact that N1s are conducted in the very patient at issue does not imply that once an N1 has been conducted the relevant set of information available for decision-making is automatically exhausted. Since relevant changes in the status of patients might occur anytime (and in fact, are likely to be happening on an ongoing basis), N1s need not provide EBM practitioners with the right probabilities. Hence, even if the probabilities provided by N1s are highly individualised and undoubtedly useful for during the clinical encounter, physicians should not assume that any further information about the patient is necessarily probabilistically irrelevant to the outcome of interest. This is a matter of judgment and close follow up, and that is precisely why I think that the estimation of the right probabilities, even when the best evidence available comes from N1s, still requires clinical Judgement.

3.5. Clinical judgment in the era of personalised and precision medicine

In this last section I shall focus on *personalised medicine* (Ginsburg and Willard, 2009; Valdes and Yin, 2016) and *precision medicine* (Cardon and Harris, 2016). These terms have been used to denote approaches to the practice of medicine that are different to Evidence Based Medicine in significant respects [64], but which deserve our attention in this chapter because they share an explicit interest in using the latest scientific technologies and research methods to provide physicians with individualised recommendations (Hamburg and Collins, 2010). My aim will be to describe to the reader how these approaches have been presented to physicians and the general public, in order to explain in what sense I think they are commendable projects to improve care, but at the same time not solutions that eliminate the PEI and so make the exercise of clinical discretion redundant.

[64] There are at least two notable differences between the EBM project and the personalised and precision medicine approach to clinical care. The first one is that given its focus on conventional clinical trials conducted for standard clinical populations (e.g. Sackett and Rosenberg, 1995; Haynes, 2002, 2006), and its connexion with health policy and the standardization of care via evidence-based guidelines and practice standards, the EBM approach, unlike precision and personalised medicine, is perceived by many authors as a project to improve healthcare that is more concerned with outcomes at a population level than with improving the outcomes of particular individuals (e.g. Saarni and Gylling, 2004; de Leon, 2012). For contrasts between the EBM approach to care and that favoured by precision and personalised medicine for specific clinical conditions, see Goldhaber, 2009; Basu 2010; Tarantini and Lanzellotti, 2010. The second important difference between EBM and personalised and precision medicine has to do with the rules of EBM (Sackett, 2000; Bluhm, 2005). While EBM has traditionally emphasised the importance of recommendations based on clinical trials, it has deemphasised the clinical relevance of basic science (Guyatt, 1991; EBM working group, 1992) including laboratory studies that focus on the study of biological properties at a cellular or molecular level with the aim of generating novel diagnostic tests and biomarkers of disease for prognostic purposes (see Baker, 2016). Nonetheless, it is worth mentioning that personalised and precision medicine share with EBM the aspiration of provide physicians with objective probabilities supported by rigorous "evidence-based" procedures (Ginsburg and Willard, 2009), and in this respect, each of these approaches distance itself from traditional clinical practice, or other approaches such as person-centred medicine, which emphasise that the right care for individuals has to consider everything the physician knows about the patient regardless of whether such information is backed up by scientifically validated data.

So, what are personalised and precision medicines? Nowadays personalised medicine and precision medicine are terms often used interchangeably (e.g. Joyner and Paneth, 2015; Jameson and Longo, 2015); this is because their supporters describe both of them as practices that use cutting-edge medical knowledge with the aim of individualising recommendations. For example, Michael Joyner and Nigel Paneth (2015) stress this aim by telling us that *"personalized or precision medicine maintains that medical care and public health will be radically transformed by prevention and treatment programs more closely targeted to the individual patient".* These writers also emphasise the role of biomarkers and technological advances by explaining that *"[Prevention and treatment] interventions will be developed by sequencing more genomes, creating bigger biobanks, and linking biological information to health data in electronic medical records (EMRs) or obtained by monitoring technologies."* (p.999).

Speaking on similar lines, Margaret Hamburg and Francis Collins (2010), highlight the auspicious prospects of personalised and precision medicine by saying that *"Researchers have discovered hundreds of genes that harbor variations contributing to human illness, identified genetic variability in patients' responses to dozens of treatments, and begun to target the molecular causes of some diseases. In addition, scientists are developing and using diagnostic tests based on genetics or other molecular mechanisms to better predict patients' responses to targeted therapy"* (p.301). And, similarly, Larry Jameson and Dan Longo (2015) emphasise that *"the convergence of genetics, informatics, and imaging, along with other technologies such as cell sorting, epigenetics, proteomics, and metabolomics, is rapidly expanding the scope of precision medicine by refining the classification of disease, often with important prognostic and treatment implications"* (p.2229).

So, as these quotes illustrate, personalised and precision medicine could be summarised as approaches to clinical care that aim at provided physicians with more refined diagnostic tests, more precise prognostic predictions, and more specific therapeutic recommendations, which can be "tailored" to particular patients by taking into account several biomarkers, in particular genetic factors (Hall et al., 2016).

The message is clear: personalised and precision medicine promise patients to assess their personal risk by means of predictive models that take into account various interactions between the patient's genomic profile and environmental variables which, it is claimed, will permit physicians to choose "the most appropriate interventions" for each patient (Hall et al, 2016).

But now, the following question arises: do these approaches truly solve the PEI and offer physicians probabilities for individuals that make the exercise of clinical judgment no longer needed or advisable?

On the one hand, one would think that the improvements generated by personalised and precision medicine are perfectly compatible with the idea of providing physicians with some latitude for the exercise of clinical judgment. After all, the fact that these approaches made possible the incorporation of new information into decision-making need not imply that place for the exercise of clinical judgment has been automatically collapsed (Dione-Labrie et al. 2010). However, not everyone has the same opinion, and advocates of EBM who tend to interpret clinical discretion as a source of unwarranted variation in healthcare due to "therapeutic illusions" (e.g. Eddy, 2005; Casarett, 2016; Howick, 2011, among others), and are keen to decrease the influence of physicians' judgment as much as possible, might be tempted to think that personalised and precision medicine provide them with a perfect case to demonstrate that true individualisation can be accomplished without resorting to clinical judgment.

Similarly, the National research council report on precision medicine conveys the idea that this approach is able to tailor medical interventions to each patient (Committee on a Framework for Developing a New Taxonomy of Disease, 2011, obtained accessed through Hunter, 2016). This again suggests that resort to clinical judgment will be less and less necessary.

Nonetheless, although I accept that personalised and precision medicine have the potential to improve clinical medicine by providing physicians with probabilities based on a larger number of factors relevant to the outcomes of interests [65], I do not think that these approaches to medical care can be used to render clinical judgment totally dispensable.

The main reason why I think this is that the kind of information provided by personalised and precision medicine (e.g. biomarkers) are only part of the story. By this I mean to say that personalised research typically incorporates only a few among many factors that might be relevant in determining the probability of a certain outcome of interest for an individual.

As I have pointed out several times throughout this thesis, there is a myriad of factors that might be relevant to the probability of the outcome of interest in any particular patient and such factors might come to be known to the physician via different sources, including her own personal clinical experience. Of course, by stressing that the factors considered by

[65] In fact, personalised and precision medicine are already providing physicians with more fine-grained probabilities in the fields of oncology and cardiovascular medicine (e.g. Gill et al., 2004; Kent et al., 2002).

personalised and precision medicine do not exhaust all the information available to the physician, I do not want to convey the idea that physicians ought not to pay attention to biomarkers. On the contrary, to the extent that biomarkers constitute additional information available to the physician at the time of recommendations, it follows from the DA approach (§ 1.5) that physicians ought to consider them if they want to estimate the right probabilities for each patient. However, it is one thing to acknowledge that the inclusion of biomarkers in clinical decisions may lead to probabilities for more refined reference classes, and quite another to claim –as the most enthusiastic supporters of personalised and precision medicine do– that these approaches can account for *"all relevant [patient's] characteristics together" (*Dorresteijn, et al., 2011) and therefore suggest that they provide the right probabilities for each individual.

Even when some recommendations are based on a biomarker that is present in the patient in question, it would be unwise to automatically assume that its causal relevance to the outcome of interest might not be subject to several interaction effects due to additional features present in the patient which can only be detected by the physician during the clinical encounter. Thus, the PEI might still arise, and to the extent that it does, physicians should be permitted to exercise their clinical judgment in their patients' best benefit.

As the physician David Hunter (2016) when discussing precision medicine says: *"In the future, we are likely to face a potentially bewildering array of probabilities — estimates of disease risk based on inherited germline sequencing and, once a disease is diagnosed, of prognosis and therapeutic options guided by "-omic" and other analyses…*" (p.713). But it is precisely because of the presence of these *"bewildering array of probabilities"* potentially applicable to the individual patient that clinical discretion is essential in allowing the physician to decide on the best recommendations for the patient.

3.6. Conclusions

This chapter was concerned with the question of whether remodelling some of the research designs and analytical methods behind EBM recommendations (in particular, conventional RCTs) could deal effectively with the PEI, and thereby provide supporters of EBM with an immediate rejoinder to my challenge to the adequacy of the EBM recommendations in the presence of additional information.

The suggestion was that suitable research improvements could account for all relevant information to the outcome of interest, and thereby collapse the distinction between the probabilities delivered by the best research available and the probabilities based on everything the physician knows about the patient.

Unfortunately for advocates of EBM, I showed that this line of argument is misguided. For a start, although *Pragmatic Randomized Controlled Trials* are extremely useful research improvements, which in my opinion can enhance the general prospects of the EBM movement, its advantage over conventional RCTs concerns the problem of external validity and in consequence does not directly address the PEI. As I pointed out, there is a sense in which, from the perspective of the treating physician, the PEI starts *after* worries about external validity have been dispelled and there is confidence that the EBM probabilities available are truly applicable to the target population. For it is at that precise moment that additional idiosyncrasies present in the patient require the application of judgment to estimate as best as possible the right probabilities for the individual.

Subgroup analyses, on the other hand, do address the PEI, albeit partially, by providing physicians with more fine-grained probabilities. This is a valuable research improvement from the point of view of individualisation, for these probabilities will as a rule be superior to the probabilities delivered by conventional analyses in the sense that they are based on more relevant information about the patient in question. Nonetheless, the advantages of subgroup analyses, even in its more sophisticated forms, are hampered by a number of practicalities that make implausible to think that they can collapse the difference between the best research probabilities and the probabilities relative to everything the physician knows about the patient. So, to put it bluntly, individualised research need not be individualised enough for many patients and in such cases clinical judgment remain necessary to arrive at the right recommendations.

As to *N of 1 Randomised Controlled Trials*, they also constitute a very useful research design, which could account for a great number of characteristics of individual patients. However, the fact that N1s are conducted in the patient in question should not be confused with the idea that the probabilities delivered by N1s exhaust all the information that might affect the outcome of interest. Even when the results of an N1 are available to the physician, the status of the patient might be subject to all sorts of variations that need to be attended to by the physician if she wants to ensure that recommendations are based on the right probabilities.

Finally, the chapter concluded with a brief discussion on the potential capabilities of *personalised medicine* and *precision medicine* to solve the PEI. Not surprisingly, even if these movements have achieved a certain level of refinement in the area of diagnostics and therapeutics, I argued that it is questionable to regard these approaches as genuine methods of individualisation, for in truth the incorporation of biomarkers is better interpreted as step forward towards a more stratified medicine (in this sense akin to product of subgroup analyses), which will permit physicians the incorporation of a greater number of relevant

factors which were previously unknown to be relevant to the outcomes of interest, but which will not eliminate the need for careful judgment if the goal is to maximize expected utility for each individual. It would be dangerous if the need to obtain prescriptions that are tailored to the individual was obscured by the possibility of basing decisions on biomarkers, on the grounds that these latter probabilities always provide the physician with the best guide to individual decisions. To repeat, in spite of technological advancements, the right probabilities remain the probabilities based on everything the physician knows about the patient.

3.7. References

Altman, D.G. (1998). Within trial variation--a false trail? *J Clin Epidemiol.* 51(4):301-3.

Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., Bernal-Delgado, E., Blomberg, N., Bock, C., Conesa, A., Del Signore, S., Delogne, C., Devilee, P., Di Meglio, A., Eijkemans, M., Flicek, P., Graf, N., Grimm, V., Guchelaar, H.J., Guo, Y.K., Gut, I.G., Hanbury, A., Hanif, S., Hilgers, R.D., Honrado, Á., Hose, D.R., Houwing-Duistermaat, J., Hubbard, T., Janacek, S.H., Karanikas, H., Kievits, T., Kohler, M., Kremer, A., Lanfear, J., Lengauer, T., Maes, E., Meert, T., Müller, W., Nickel, D., Oledzki, P., Pedersen, B., Petkovic, M., Pliakos, K., Rattray, M., I Màs, J.R., Schneider, R., Sengstag, T., Serra-Picamal, X., Spek, W., Vaas, L.A., van Batenburg, O., Vandelaer, M., Varnai, P., Villoslada, P., Vizcaíno, J.A., Wubbe, J.P. and Zanetti, G. (2016). Making sense of big data in health research: Towards an EU action plan. *Genome Med*. 8(1):71.

Alyass, A., Turcotte, M. and Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics*. 8:33.

Baker, M. (2016). Increasing precision in medicine - tackling the bottleneck of variant interpretation. *Drugs Today (Barc).* 52(7):395-8.

Barlow, D.H. and Hersen, M. (1984). *Single Case Experimental Designs: Strategies for Studying Behavior Change.* 2nd Ed. New York, Pergamon.

Basu, S. (2010). Personalized versus evidence-based medicine with PET-based imaging. *Nat Rev Clin Oncol.* 7(11):665-8.

Beebee, H., Hitchcock, C. and Menzies, P. (2010). *Oxford Handbook of Causation.* Oxford, Oxford University Press.

Berry, D.A. (2006). Bayesian clinical trials. *Nat Rev Drug Discov.* 5(1):27-36.

Bleeker, S.E., Moll, H.A., Steyerberg, E.W., Donders, A.R., Derksen-Lubsen, G., Grobbee, D.E. and Moons, K.G. (2003). External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol.* 56(9):826-32.

Bluhm, R. (2005). From hierarchy to network: a richer view of evidence for evidence-based medicine. *Perspect Biol Med.* 48(4):535-47.

Borenstein, M., Hedges, L.V., Higgins, J.P. and Rothstein, H.R. (2009). *Critics of Meta-analysis.* In: *Introduction to meta-analysis.* Chichester, John Wiley & Sons Ltd. pp. 325-6.

Bothwell, L.E. and Podolsky, S.H. (2016). The Emergence of the Randomized, Controlled Trial. *N Engl J Med.* 375(6):501-4.

Brookes, S.T., Whitley, E., Peters, T.J., Mulheran, P.A., Egger, M. and Davey Smith, G. (2001). Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess.* 5(33):1-56.

Brookes, S.T., Whitely, E., Egger, M., Smith, G.D., Mulheran, P.A. and Peters, T.J. (2004). Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol.* 57(3):229-36.

Cardon, L.R. and Harris, T. (2016). Precision medicine, genomics and drug discovery. *Hum Mol Genet*. pii: ddw246.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement.* Oxford, Clarendon Press.

Cartwright, N. (2007). Are RCTs the Gold Standard? *BioSocieties*. 2:11-20

Cartwright, N. (2009). *What is this thing called efficacy.* In *Philosophy of the social sciences: Philosophical theory and scientific practice*. C. Mantzavinos. Cambridge, Cambridge University Press. pp. 185-206.

Cartwright, N. (2010). Will this policy work for you? Predicting effectiveness better: How philosophy helps. *Philosophy of Science.* 79(5):973-89.

Cartwright, N. (2013). Knowing what we are talking about: Why evidence doesn't always travel. *Evidence & Policy: A Journal of Research, Debate and Practice.* 9(1):97-112.

Casarett, D. (2016). The Science of Choosing Wisely--Overcoming the Therapeutic Illusion. *N Engl J Med.* 374(13):1203-5.

Chalkidou, K., Tunis, S., Whicher, D., Fowler, R. and Zwarenstein, M. (2012).The role for pragmatic randomized controlled trials (pRCTs) in comparative effectiveness research. *Clin Trials.* 9(4):436-46.

Christmas, D. (2014). Has the pharmaceutical industry commandeered evidence-based medicine? *Scottish Universities Medical Journal*. 3(supp1):s12- s18.

Clarke, B., Gillies, D., Illari, P., Russo, F. and Williamson, J. (2014). Mechanisms and the Evidence Hierarchy. *Topoi.* 33:339–60.

Collier, J., Longmore, M., Amarakone, K. (2013). Oxford Handbook of Clinical Specialties (Oxford Medical Handbooks). New York, Oxford University Press.

Collins, G.S., de Groot, J.A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L.M., Moons, K.G. and Altman, D.G. (2014). External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol.* 14:40. doi: 10.1186/1471-2288-14-40.

de Leon, J. (2012). Evidence-Based Medicine versus Personalized Medicine: Are They Enemies? *J Clin Psychopharmacol.* 32(2):153-64.

Dion-Labrie, M., Fortin, M.C., Hébert, M.J. and Doucet, H. (2010). The use of personalized medicine for patient selection for renal transplantation: physicians' views on the clinical and ethical implications. *BMC Med Ethics.*11:5.

Doll, R. (1992). Sir Austin Bradford Hill and the progress of medical science. *BMJ.* 305(6868):1521-6.

Dorresteijn, J.A., Visseren, F.L., Ridker, P.M., Wassink, A.M., Paynter, N.P., Steyerberg, E.W., van der Graaf, Y. and Cook, N.R. (2011). Estimating treatment effects for individual patients based on the results of randomised clinical trials. *BMJ.* 343:d5888.

EBM Working Group. (1992). Evidence-Based Medicine: a new approach to teaching the practice of medicine. *JAMA.* 268(17):2420-5.

Eddy, D.M. (2005). Evidence-based medicine: a unified approach. *Health Aff (Millwood).* 24(1):9-17.

Eells, E. (1987). Probabilistic Causality: Reply to John Dupre. *Philosophy of Science.* 54:105-14.

Feinstein, A.R. (1967). *Clinical Judgment.* Baltimore, Williams & Wilkins Co.

Feinstein, A.R. (1968). Clinical epidemiology. I. The populational experiments of nature and of man in human illness. *Ann Intern Med.* 69(4):807-20.

Feinstein, A.R. (1998). The Problem of Cogent Subgroups: A Clinicostatistical Tragedy. *J Clin Epidemiol.* 51(4):297-9.

Fisher, R.A. (1990). *The design of experiments.* In *Statistical Methods, Experimental Design and Scientific Inference.* J.H. Bennett (ed). Oxford, Oxford University Press.

Fortin, M., Dionne, J., Pinho, G., Gignac, J., Almirall, J. and Lapointe, L. (2006). Randomized Controlled Trials: Do They Have External Validity for Patients With Multiple Comorbidities? *Ann Fam Med.* 4(2):104-8.

Fuller, J. (2016). *The new medical model: chronic disease and evidence-based medicine.* PhD Thesis, University of Toronto.

Fuller, J. and Flores, L.J. (2015). The Risk GP Model: the standard model of prediction in medicine. *Stud Hist Philos Biol Biomed Sci.*54:49-61.

Fuller, J. and Flores, L.J. (2016). Translating Trial Results in Clinical Practice: the Risk GP Model. *J Cardiovasc Transl Res.* 9(3):167-8.

Gabler, N.B., Duan, N., Liao, D., Elmore, J.G., Ganiats, T.G. and Kravitz, R.L. (2009). Dealing with heterogeneity of treatment effects: is the literature up to the challenge? *Trials.* 10:43.

Gabler, N.B., Duan, N., Vohra, S. and Kravitz, R.L. (2011). N-of-1 trials in the medical literature: a systematic review. *Med Care.* 49:761-8.

Gill, S., Loprinzi, C.L., Sargent, D.J., Thome, S.D., Alberts, S.R., Haller, D.G., Benedetti, J., Francini, G., Shepherd, L.E., Francois Seitz, J., Labianca, R., Chen, W., Cha, S.S., Heldebrant, M.P. and Goldberg, R.M. (2004). Pooled analysis of fluorouracil-based adjuvant therapy for stage II and III colon cancer: who benefits and by how much? *J Clin Oncol*. 22(10):1797-806.

Ginsburg, G.S. and Willard, H.F. (2009). Genomic and personalized medicine: foundations and applications. *Transl Res.* 154(6):277-87.

Goldhaber, S.Z. (2009). Optimal duration of anticoagulation after venous thromboembolism: fixed and evidence-based, or flexible and personalized? *Ann Intern Med.* 150(9):644-6.

Godwin, M., Ruhland, L., Casson, I., MacDonald, S., Delva, D., Birtwhistle, R., Lam, M. and Seguin, R. (2003). Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. *BMC Med Res Methodol.* 3:28.

Greenhalgh, T., Howick, J. and Maskrey, N.; Evidence Based Medicine Renaissance Group. (2014). Evidence based medicine: a movement in crisis? *BMJ.* 348:g3725.

Guyatt, G., Sackett, D., Taylor, D.W., Chong, J., Roberts, R. and Pugsley, S. (1986). Determining optimal therapy--randomized trials in individual patients. *N Engl J Med*. 314(14):889-92.

Guyatt, G., Sackett, D., Adachi, J., Roberts, R., Chong, J., Rosenbloom, D. and Keller, J. (1988). A clinician's guide for conducting randomized trials in individual patients. CMAJ. 139(6):497-503.

Guyatt, G.H., Heyting, A., Jaeschke, R., Keller, J., Adachi, J.D. and Roberts, R.S. (1990). N of 1 randomized trials for investigating new drugs. *Control Clin Trials*. 11(2):88-100.

Guyatt, G.H. (1991). Evidence-based medicine. *ACP J Club*. 114:A16.

Guyatt, G.H., Haynes, R.B., Jaeschke, R.Z., Cook, D.J., Green, L., Naylor, C.D., Wilson, M.C. and Richardson, W.S. (2000). Users' Guides to the Medical Literature. XXV. Evidence-Based Medicine: Principles for Applying the Users' Guides to Patient Care. *JAMA.* 284(10):1290-6.

Hacking, I. (1988). Telepathy: Origins of Randomization in Experimental Design. *Isis.* 79(3):427-51.

Hacking, I. (2001). *An introduction to probability and inductive logic.* Cambridge, Cambridge University Press.

Hall, M.A., Moore, J.H. and Ritchie, M.D. (2016). Embracing Complex Associations in Common Traits: Critical Considerations for Precision Medicine. *Trends Genet.* 32(8):470-84.

Hamburg, M.A. and Collins, F.S. (2010). The path to personalized medicine. *N Engl J Med*. 363:301-4.

Haynes, R.B. (2002). What kind of evidence is it that Evidence-Based Medicine advocates want health care providers and consumers to pay attention to? *BMC Health Services Research*. 2:3.

Haynes, R.B. (2006). *Clinical epidemiology: how to do clinical practice research*. 3rd ed. Philadelphia, Lippincott Williams & Wilkins.

Hayward, R.A., Kent, D.M., Vijan, S. and Hofer, T.P. (2006). Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Med Res Methodol.* 6:18.

Hemingway, H., Croft, P., Perel, P., Hayden, J.A., Abrams, K., Timmis, A., Briggs, A., Udumyan, R., Moons, K.G., Steyerberg, E.W., Roberts, I., Schroter, S., Altman, D.G. and Riley, R.D.; PROGRESS Group. (2013). Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ.* 346:e5595.

Hernán, M.A. and Robins, J.M. (2006). Estimating causal effects from epidemiological data. *J Epidemiol Community Health.* 60(7): 578–86.

Hill, A.B. (1952). The clinical trial. *N Engl J Med.* 247:113-9.

Hitchcock, C. (2010). *"Probabilistic Causation, The Stanford Encyclopedia of Philosophy"* Retrieved 06 May, 2015, from http://plato.stanford.edu/archives/win2012/entries/causation-probabilistic/

Hotopf, M. (2002). The pragmatic randomised controlled trial. *Adv Psychiatr Treat.* 8:326-33.

Howitz, J. (2011). *The Philosophy of Evidence-based Medicine.* Oxford, Wiley-Blackwell.

Howick, J., Glasziou, P. and Aronson, J.K. (2013). Problems with using mechanisms to solve the problem of extrapolation. *Theor Med Bioeth.* 34(4):275-91.

Howson, C. and Urbach, P. (1989). *Scientific Reasoning: the Bayesian Approach.* La Salle, Open Court.

Hunter, D.J. (2016). Uncertainty in the Era of Precision Medicine. *N Engl J Med.* 375(8):711-3.

Ioannidis, J.P. (2016). Evidence-based medicine has been hijacked: a report to David Sackett. *J Clin Epidemiol.* 73:82-6.

Jameson J.L. and Longo D.L. (2015). Precision medicine-personalized, problematic, and promising. *N Engl J Med.* 372(23):2229-34.

Johnson, K.E., Tachibana, C., Coronado, G.D., Dember, L.M., Glasgow, R.E., Huang, S.S., Martin, P.J., Richards, J., Rosenthal, G., Septimus, E., Simon, G.E., Solberg, L., Suls, J., Thompson, E. and Larson, E.B. (2014). A guide to research partnerships for pragmatic clinical trials. *BMJ.* 349:g6826.

Keene, O.N. and Garret, A.D. (2014). Subgroups: Time to Go Back to Basic Statistical Principles? *J Biopharm Stat.* 24(1):58-71.

Jones, H.E., Ohlssen, D.I., Neuenschwander, B., Racine, A. and Branson, M. (2011). Bayesian models for subgroup analysis in clinical trials. *Clin Trials.* 8(2):129-43.

Joyner, M.J. and Paneth, N. (2015). Seven Questions for Personalized Medicine. *JAMA.* 314(10):999-1000.

Kent, D.M., Hayward, R.A., Griffith, J.L., Vijan, S., Beshansky, J.R., Califf, R.M. and Selker, H.P. (2002). An independently derived and validated predictive model for selecting patients with myocardial infarction who are likely to benefit from tissue plasminogen activator compared with streptokinase. *Am J Med.* 113(2):104-11.

Kent, D.M. and Hayward, R.A. (2007a). Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA.* 298:1209-12.

Kent, D.M. and Hayward, R.A. (2007b). When Averages Hide Individual Differences in Clinical Trials: Analyzing the results of clinical trials to expose individual patient's risks might help doctors make better treatment decisions. *American Scientist*. 95(1):60-8.

Kent, D.M., Rothwell, P.M., Ioannidis, J.P., Altman, D.G. and Hayward, R.A. (2010). Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials.* 11:85.

Khorsan, R., and Crawford, C. (2014). How to Assess the External Validity and Model Validity of Therapeutic Trials: A Conceptual Approach to Systematic Review Methodology. *Evid Based Complement Alternat Med.* 2014:694804.

Kravitz, R.L., Duan, N., Niedzinski, E.J., Hay, M.C., Subramanian, S.K. and Weisner, T.S. (2008). What ever happened to N-of-1 trials? Insiders' perspectives and a look to the future. *Milbank Q.* 86(4):533-55.

Kravitz, R.L. Duan, N. (eds.) and the DEcIDE Methods Center N-of-1 Guidance Panel (2014). *Design and Implementation of N-of-1 Trials: A User's Guide.* AHRQ Publication No. 13(14)-EHC122-EF. Rockville, Agency for Healthcare Research and Quality.

Kumar, D. (2011). The personalised medicine. A paradigm of evidence-based medicine. *Ann Ist Super Sanita.* 47(1):31-40.

Lilford, R.J. and Jackson, J. (1995) Equipoise and the ethics of randomization. *J R Soc Med*. 88:552-9

Longford, N.T. and Nelder, J.A. (1999). Statistics versus statistical science in the regulatory process. *Statist Med.* 18:2311-20.

MacPherson, H. (2004). Pragmatic clinical trials. *Complementary Therapies in Medicine*. 12:136-40.

Mannucci, P.M. and Nobili, A; REPOSI Investigators. (2014). Multimorbidity and polypharmacy in the elderly: lessons from REPOSI. *Intern Emerg Med.* 9(7):723-34.

Marks, H.M. (1997). *The Progress of Experiment, Science and Therapeutic Reform in the United States, 1900–1990.* Cambridge, Cambridge University Press.

Marks, D.M., J, T. and Pae, C.U. (2009). Innovations in clinical research design and conduct in psychiatry: shifting to pragmatic approaches. *Psychiatry Investig.* 6(1):1-6.

Matthews, J.R. (1995). *Quantification and the Quest for Medical Certainty*. Princeton, Princeton University Press.

Mayo, D.G. and Spanos, A. (2010). *Error and Inference: Recent Exchanges on Experimental Reasoning, reliability and the Objectivity and Rationality of Science.* Cambridge, Cambridge University Press.

Naggara, O., Raymond, J., Guilbert, F. and Altman, D.G. (2011). The problem of subgroup analyses: an example from a trial on ruptured intracranial aneurysms. *AJNR Am J Neuroradiol.* 32(4):633-6.

Oxman, A.D. and Guyatt, G.H. (1992). A consumer's guide to subgroup analyses. *Ann Intern Med.* 116(1):78-84.

Papineau, D. (1985). Probabilities and Causes. *Journal of Philosophy*. 82:57-74.

Papineau, D. (1989). *Pure. Mixed, and Spurious Probabilities and their Significance for a Reductionist Theory of Causation.* In *Minnesota Studies in the Philosophy of Science XIII: Scientific Explanation.* P. Kitcher and W. Salmon (eds). Minneapolis, University of Minnesota Press. pp. 307-48.

Papineau, D. (1994). The Virtues of Randomization. *Brit J Phil Sci*. 45:437-50.

Papineau, D. (2012). *Correlations and causes.* In *Philosophical Devices: Proofs, Probabilities, Possibilities, and Sets.* Oxford, Oxford University Press. Chapter 9. pp. 119-33.

Parekh, A.K., Goodman, R.A., Gordon, C. and Koh, H.K.; HHS Interagency Workgroup on Multiple Chronic Conditions. (2011). Managing multiple chronic conditions: a strategic framework for improving health outcomes and quality of life. *Public Health Rep.* 126(4):460-71.

Patsopoulos, N.A. (2011). A pragmatic view on pragmatic trials. *Dialogues Clin Neurosci.* 13(2):217-24.

Pearce, W., Raman, S. and Turner, A. (2015). Randomised trials in context: practical problems and social aspects of evidence-based medicine and policy. *Trials.* 16:394.

Petersen, M.K., Andersen, K.V., Andersen, N.T. and Søballe, K. (2007). "To whom do the results of this trial apply?" External validity of a randomized controlled trial involving 130 patients scheduled for primary total hip replacement. *Acta Orthop.* 78(1):12-8.

Peto, R., Collins, R. and Gray, R. (1995). Large-scale randomized evidence: large, simple trials and overviews of trials. *J Clin Epidemiol*. 48(1):23-40.

Peto, R. (2011). Current misconception 3: that subgroup-specific trial mortality results often provide a good basis for individualising patient care. *Br J Cancer.* 104:1057-8.

Pocock, S.J., Assmann, S.E., Enos, L.E. and Kasten, L.E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med*. 21(19):2917-30.

Porter, R. (1997). *The Greatest Benefit to Mankind: A Medical History of Humanity from Antiquity to the Present.* London, W. W. Norton & Company.

Riley, R.D., Hayden, J.A., Steyerberg, E.W., Moons, K.G., Abrams, K., Kyzas, P.A., Malats, N., Briggs, A., Schroter, S., Altman, D.G. and Hemingway, H.; PROGRESS Group. (2013). Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med.* 10(2):e1001380.

Rothwell, P.M. (1995). Can overall results of clinical trials be applied to all patients? *Lancet*. 345:1616-9.

Rothwell, P.M. (2005a). External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet.* 365:82-93.

Rothwell, P.M. (2005b). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet.* 365:176-86.

Rothwell, P.M. (2006). Factors That Can Affect the External Validity of Randomised Controlled Trials. *PLoS Clin Trials.* 1(1): e9.

Rothwell, P.M. (2010). Commentary: External validity of results of randomized trials: disentangling a complex concept. *Int J Epidemiol.* 39(1):94-6.

Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International studies in the philosophy of science.* 21(2):157-70.

Russo, F. and Williamson, J. (2011). Epistemic causality and evidence-based medicine. *Hist Philos Life Sci.* 33(4):563-81.

Saarni, S.I. and Gylling, H.A. (2004). Evidence based medicine guidelines: a solution to rationing or politics disguised as science? J *Med Ethics.* 30(2):171-5.

Sackett, D.L. and Rosenberg, W.M. (1995). On the need for evidence-based medicine. *J Public Health Med.* 17(3):330-4.

Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B. and Richardson, W.S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ*. 312(7023):71-2.

Sackett, D.L. (2000). The fall of "clinical research" and the rise of "clinical-practice research". *Clin Invest Med.* 23(6):379-81.

Sculpher, M. (2010). Reflecting heterogeneity in patient benefits: the role of subgroup analysis with comparative effectiveness. *Value Health.* 13(Suppl 1):S18-21.

Senior, H.E., Mitchell, G.K., Nikles, J., Carmont, S.A., Schluter, P.J., Currow, D.C., Vora, R., Yelland, M.J., Agar, M., Good, P.D. and, Hardy, J.R. (2013). Using aggregated single patient (N-of-1) trials to determine the effectiveness of psychostimulants to reduce fatigue in advanced cancer patients: a rationale and protocol. *BMC Palliat Care.* 12(1):17.

Senn, S. and Harrell, F. (1997). On wisdom after the event. *J Clin Epidemiol.* 50(7):749-51.

Sharma, S. (2014). Nanotheranostics in evidence based personalized medicine. *Curr Drug Targets.* 15(10):915-30.

Shaw, E.C. and Johnson, P.W. (2012). Stratified medicine for cancer therapy. *Drug Discov Today*. 17(5-6):261-8.

Steckler, A. and McLeroy, K.R. (2008). The Importance of External Validity. *Am J Public Health.* 98(1):9-10.

Steeples, L., and Mercieca, K. (2012). Acute conjunctivitis in primary care: antibiotics and placebo associated with small increase in the proportion cured by 7 days compared with no treatment. *Evid Based Med.* 17(6):177-8.

Steyerberg, E.W. (2009). *Clinical prediction models: a practical approach to development, validation, and updating.* New York, Springer.

Sun, X., Briel, M., Busse, J.W., You, J.J., Akl, E.A., Mejza, F., Bala, M.M., Bassler, D., Mertz, D., Diaz-Granados, N., Vandvik, P.O., Malaga, G., Srinathan, S.K., Dahm, P., Johnston, B.C., Alonso-Coello, P., Hassouneh, B., Walter, S.D., Heels-Ansdell, D., Bhatnagar, N., Altman, D.G. and Guyatt, G.H. (2012). Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ.* 344:e1553.

Sun, X., Ioannidis, J.P., Agoritsas, T., Alba, A.C. and Guyatt, G. (2014). How to use a subgroup analysis: users' guide to the medical literature. *JAMA.* 311(4):405-11.

Tarantini, G. and Lanzellotti, D. (2010). Three-vessel coronary disease in diabetics: personalized versus evidence-based revascularization strategy. *Future Cardiol*. 6(6):797-809.

Travers, J., Marsh, S., Caldwell, B., Williams, M., Aldington, S., Weatherall, M., Shirtcliffe, P. and Beasley, R. (2007). External validity of randomized controlled trials in COPD. *Respir Med.* 101(6):1313-20.

Treeweek, S. and Zwarenstein, M. (2009). Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials.* 10:37.

Uijen, A.A., Bakx, J.C., Mokkink, H.G. and van Weel, C. (2007). Hypertension patients participating in trials differ in many aspects from patients treated in general practices. *J Clin Epidemiol.* 60(4):330-5.

Valdes, R. Jr and Yin, D.T. (2016). Fundamentals of Pharmacogenetics in Personalized, Precision Medicine. *Clin Lab Med.* 36(3):447-59.

Vandenbroucke, J.P., Broadbent, A. and Pearce, N. (2016). Causality and causal inference in epidemiology: the need for a pluralistic approach. *Int J Epidemiol.* pii: dyv341.

Vist, G.E., Bryant, D., Somerville, L., Birminghem, T. and Oxman, A.D. (2008). Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate. *Cochrane Database of Systematic Reviews*. (3):MR000009.

Vohra, S., Shamseer, L., Sampson, M., Bukutu, C., Schmid, C.H., Tate, R., Nikles, J., Zucker, D.R., Kravitz, R., Guyatt, G., Altman, D.G. and Moher, D.; CENT Group. (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. *BMJ*. 350:h1738.

Wang, R., Lagakos, S.W., Ware, J.H., Hunter, D.J. and Drazen, J.M. (2007). Statistics in medicine--reporting of subgroup analyses in clinical trials. *N Engl J Med*. 357(21):2189-94.

Worrall, J. (2007). Why There's No Cause to Randomize. *Br J Philos Sci*. 58(3):451-88.

Worrall, J. (2011). Causality in medicine: getting back to the Hill top. *Prev Med*. 53(4-5):235-8.

Ziebach, R., Pietsch-Breitfeld, B., Bichler, M., Busch, A., Riethmüller, J. and Stern, M. (2001). Bronchodilatory effects of salbutamol, ipratropium bromide, and their combination: double-blind, placebo-controlled crossover study in cystic fibrosis. *Pediatr Pulmonol*. 31(6):431-5.

Chapter 4: The DA: counterpoints and further objections

4.1. Abstract

This chapter elaborates on the Discretionary Approach (DA). I begin by distinguishing it from two standpoints that might be associated with it: Medical Particularism and Person-Centred Medicine. Then I address a range of possible criticisms of the DA: a lack of normative novelty; a lack of descriptive novelty; the charge of truistic obviousness; and the charge of impracticality. Finally, the chapter closes with a discussion of a more fundamental objection, namely: that the implementation of the DA may empirically worsen clinical care. My analysis indicates that while some concerns about the practical consequences of the DA are reasonable, EBM's current empirical case again the DA is questionable. I conclude that updated empirical evidence is required to address the empirical effectiveness of EBM versus the DA—evidence that will then be provide in the final two chapters of the thesis.

4.2. Useful counterpoints

4.2.1. Medical particularism

One of the reasons why some readers, in particular those sympathetic with EBM, might be highly suspicious of a proposal along the lines of the DA is because it could be read as an instance of a theoretical position I shall denote "Medical Particularism" (MP), which stands in sharp opposition to the principles of EBM care. In this section my argument will be that, on first inspection, the DA may appear similar to such stance, but closer examination reveals that the DA is in fact more akin to EBM care than to MP.

Let me begin by offering a brief description of MP, before turning to its alleged connexion with the DA.

What exactly is MP? This term is used to describe a theoretical position akin to that defended by the philosopher Jonathan Dancy in ethical theory (Dancy, 2004). It is important to note immediately that, as far as the author knows, MP is not a standpoint that has been officially defended in print by medical scholars or philosophers of medicine. Rather, MP is a position that has been mentioned in workshops, seminars, and critical discussions on the problems of the EBM approach (e.g. a number of workshops organised by the Centre for the Humanities and Health at King's College London).

A useful way to imagine an advocate of MP is as someone who believes that optimal clinical care does not follow from the application of rules to categories of patients (no matter how fine-grained). According to MP, the care of individual patients always rests on a special kind of judgment or intuition that considers the patient as something more than a set of features. Thus, understandably, physicians defending MP would presumably object to a project like EBM, which is implicitly premised on the applicability of general efficacy claims to particular patients, for from the standpoint of MP there is an insurmountable schism between two types of medical knowledge: general and particular.

Consider, as an illustration, the remarks of the philosopher Jonathan Wolff at a conference on the topic: *"a medical particularist would not assume that everything fits neatly together – that there is or could be a manual for the human body – and will have to proceed in absence of such a rulebook. Thus, [the medical particularist] would need far more than a range of medical principles, rules, regularities or research findings and the ability to apply them to be a good medical judge."* (Bullock and Kingma, 2014, p.999).

The aforementioned quote illustrates that MP seems to raise a fundamental objection against the direct applicability of research evidence to individuals. So, from this perspective,

supporters of EBM are just missing the point, not because research (even in its more sophisticated forms) cannot account for a sufficiently great number of relevant features present in the individual, but because the application of *general "population-level" knowledge* to a *particular patient* necessarily requires some form of judgment or intuition that considers the patient in all its individuality, by which the MP means much more than a set of features, however detailed.

Now, why would someone associate an approach such as the DA with MP? The most obvious answer to this question is that both the DA and MP raise objections against the EBM approach, and that the objections raised have something to do with the importance of clinical judgment.

However, while the DA criticises the automatic extrapolation of research findings to individuals, even in the presence of extra information available to the physician, MP objects something far more fundamental: EBM is in the wrong track with respect to clinical inference at the most basic level, for to assume that research can –even in principle– be readily applicable to individuals is completely misguided from the point of view of MP. So, while a physician that supports the DA would not comply about recommendations that ignore the PEI, she would not deny the idea that research can be applicable to individuals. By contrast, the physician advocating MP would take issue with such application even if she has a research probability for a reference class defined by everything she knows about the patient.

In addition, while the DA only requests more room for the exercise of judgment to account for the features not considered by the best research available (that is, to address the PEI), MP asserts that the gap between the general and the particular can only be bridged by some type judgment which is not clearly specified.[66]

Recall the point expressed in chapter 1 (footnote 21). If a physician who supports the DA happens to have access to certain probabilities for male smokers of a certain age, and the only thing she knows about a particular patient is that she is a male smoker of the corresponding age, then she would not have problems in applying this kind of general medical knowledge to his patient. This situation illustrates that, from the standpoint of the DA, there is no ontological division between the general and the particular, and therefore there is no gap to be bridged using some kind of intuition. From the standpoint of the DA, the EBM approach is wrong in neglecting the PEI: because in real clinical scenarios there are many situations in

[66] More specifically the difference between the DA and MP is the role assigned to clinical judgment. While in the DA it is explicit and circumscribed to address the PEI, the role assigned to judgment by MP leaves much more room for interpretation. In my view, apart from the general idea that judgment is necessary to bridge the gap between the general and the particular, it remains obscure what is involved in such kind of judgment or intuition.

which the physician happens to know more about the patient than she has valid research data for, EBM requires general revision so as to provide physicians with more room for clinical judgment.[67]

Thus, while physicians who exercise their judgment in accordance with the DA would be happy to accept probabilities from research when they were probabilities obtained in the reference class conformed by everything they know about the patient in question; a physician who embraces MP, on the other hand, would reject any research probabilities, regardless of the reference class in which these were obtained, because according to MP, there is no such thing as the right reference class: clinical recommendations should not be understood in terms of reference classes, because clinical recommendations for individuals transcend anything written as a general rule such as "in circumstances a, b, c…n, the physician should do x".

So, in conclusion, close attention to the claims endorsed by the DA and MP, makes it clear that, while both approaches raise concerns about clinical judgment and the application of research evidence, these two approaches do it in different terms and with different implications. Hence, there is a clear sense in which the DA is much closer to EBM than to MP.

4.2.2. Person-centred medicine

Person-centred medicine (PCM) is another approach to clinical care that, in one of its current forms, emerged in response to the EBM movement. This approach has also some similarities with the DA, so it is worth clarifying what distinguishes the two. My main point will be that while the main criticism of the DA to EBM targets the PEI, the main objection of PCM is that EBM has been focused on providing research findings for the clinical management of *"diseases"* rather than evidence relevant to *"the person"* as whole (Miles and Mezzich, 2011).

PCM, unlike MP, is an approach that has been defended in the medical literature, with its own journals[68] as well as associations[69] supporting it. Nonetheless, despite its popularity in certain circles of physicians, mostly opposed to the EBM approach, PCM has been vaguely and

[67] This point could be stressed even more emphatically by saying that from the perspective of the DA, when the physician is commit herself to the claim *"this treatment works for Mr Smith" she* is also implicitly committing herself to the general claim *"this treatment works for patients of the kind I know Mr Smith to be"*. Furthermore, since the DA does not draw a an ontological line between the general and the particular, this approach is less subject to charges of "anti-scientism", for there is no sense in which the DA discourages the generation of more and better scientific research, which could help physicians to improve the care of individuals.

[68] The two most famous are the "International Journal of Person Centered Medicine" and "The European Journal for Person Centered Healthcare".

[69] For example, The International College of Person-centered Medicine

variously defined as a standpoint for clinical care (Mezzich et al. 2010, 2011; Miles and Loughlin, 2011; Roberti di sarsina et al. 2010, 2012).

Consider the following descriptions of PCM as illustrations:

The International College of Person Centred Medicine (ICPCM) defined it as an approach *"dedicated to the promotion of health as a state of physical, mental, socio-cultural and spiritual wellbeing as well as to the reduction of disease, and founded on mutual respect for the dignity and responsibility of each individual person."* (ICPCM, 2016).

Mezzich and colleagues (2010) conceive it as: *"a medicine of the person (of the totality of the person's health, including its ill and positive aspects), for the person (promoting the fulfilment of the person's life project), by the person (with clinicians extending themselves as full human beings, well grounded in science and with high ethical aspirations) and with the person (working respectfully, in collaboration and in an empowering manner through a partnership of patient, family and clinicians)"* (Quoted from Miles and Loughlin, 2011.p.533).

Other authors, such as Roberti di Sarsina and colleagues (2012), have defined PCM a little more precisely as an approach to care which aims at *"individualizing treatments beyond clinical guidelines to suit the whole person in the context of his or her bio-psycho-spiritual biography." (*p.1*)*, but they have also included it within its boundaries complementary medicine by referring to PCT as *"a synergistic and harmonious blend of conventional and complementary medicine, but looks open to future developments."* (Roberti Di Sarsina and Iseppato, 2010, p.277).[70]

The main idea behind these descriptions has been summarized by the founders of this movement, Andrew Miles and Juan Mezzich, when they say that PCM stresses that clinical care should be less "disease centred" and should include socio-cultural factors in decision-making (Miles and Mezzich, 2011). In this respect, supporters of PCM have criticised EBM by arguing that knowledge of diseases *"cannot but fail to understand the essence of the human person and the totality of 'what is wrong'"(*Miles, 2012, p.329). In their view, there is tension between the *"disease centeredness"* of EBM and the focus on the "person" which is characteristic of PCM (Loughlin, 2014).[71]

[70] Notice that, advocates of PCM, in some passages, get very close to the idea of MP. Consider as an illustration this quote from Roberti di Sardinia and colleagues (2012): *"Personalisation in medicine needs to consider the individual as a unique being where the whole is more than the sum of its parts."*(p. 9), which seems to suggest that there is an insurmountable gap between research and the individual, and no just an insufficient level of detail in the reference classes for which we have valid research data.

[71] In the opinion of some supporters of PCM, the difference between the kind of evidence provided by conventional clinical trials, which tells us what works for general classes of patients, and the kind of evidence

Given this description of PCM as an approach concerned with the "person", there is some superficial similarity between the DA and PCM. However, this similarity depends on how one understands concepts such as "disease" and "person". As I have emphasised through this dissertation, the main objection of the DA to the EBM approach has to do with the PEI. At its simplest, this is the problem that the physician often knows more about the patient than she has valid research data for (§ 1.4.4). If we assume that the use of the concept "person" in the context of PCM is equivalent to asking for a more fine-grained reference class of patients, then one could infer that supporters of PCM are criticizing EBM on the same grounds as the DA. For, on this reading, PCM advocates are objecting that the evidence usually provided by EBM is not necessarily applicable to the patient, since this evidence leaves too many relevant features out of the equation. Put it in terms of probabilities, one could say that PCM would agree that EBM probabilities are not the right probabilities for the patient, because they were not obtained in the right reference class.

But all this is arguably assuming too much. Since there is no mention of probabilities in the discourse of PCM, it might well be that PCM only wants to stress that EBM recommendations have ignored patients' preferences and values, and are in no way concerned with the adequacy of EBM probabilities.

In fact, apart from notable exceptions (e.g. Miles and Loughlin, 2011), most authors in the PCM camp object that EBM's recommendations neglect the patient's interest and, therefore, are focused on the control of disease rather than on what truly matters to the patient (ICPCM, 2016; Mezzich et al. 2010, 2011, Roberti di Sarsina et al. 2012).[72]

In truth, the worry raised by the DA about the EBM approach is much more focused than the broad concerns expressed by advocates of PCM. It is one thing to object that EBM supports recommendations that ignore additional features about the patient that has not been considered from research probabilities; it is another to object that the EBM approach provides findings applicable to diseases that are irrelevant to a particular patient in all her individuality.

Likewise, with respect to clinical judgment, the DA specifies a much more focused role than that urged by PCM. While the main task of the physician who embraces the DA is to address the PEI by considering everything she knows about the patient to estimate the right probabilities, the role of judgment within the context of PCM remains unclear. So, although

---

needed to fulfil the aim of PCM, is so vast that talk of an "evidence-based person centred medicine should be read as an oxymoron" (Kaltoft et al., 2011). I see no reason to accept this, given that it is perfectly possible to imagine a form of clinical care informed by research evidence, as EBM recommends, but focused on the patient, as PCM advises.

[72] Notice that while the DA explicitly requires physicians to pay attention to research evidence, PCM is less explicit about the importance of such evidence for clinical decision-making taken to the extreme, PCM seems prepared to neglect research evidence on similar grounds to MP.

both these approaches seem to be concerned with the fact that research findings do not tell physicians all they need to know to arrive at the correct recommendations, they do it in very different terms, using different language, and it is unclear that the proposed solutions would lead to the same recommendations.

4.3. Further objections

4.3.1. A normative "old hat"

This objection might target two different aspects of the DA: (i) the emphasis on the "right probabilities" for clinical recommendations, and (ii) the role ascribed to clinical judgment.

Let us consider the idea of the "the right probabilities" for clinical recommendations first. In this regard, Jeremy Howick, in his "The Philosophy of EBM" (2011), suggests that the EBM approach is committed with the view that *"all relevant evidence must be considered"*, and therefore indicates that this movement endorses the familiar *"principle of total evidence"(PTE)* (p.15). If Jeremy Howick is right that EBM is truly committed to the principle of total evidence, then there would indeed be nothing novel or distinctive in the DA's suggestion that "*the right probabilities for clinical recommendations are the probabilities in the reference class defined by everything the physician knows about the patient".*

However, this line of reasoning is misguided. When Howick says that EBM endorses the PTE, he means something rather different from the adoption of the DA. (After all, as we saw in Chapter 2, the 4S Model and the EBGs model certainly do not support the idea that EBM accepts DA probabilities as the right probabilities for clinical recommendations.)

On closer examination, Howick's advocacy of PTE turns out to be based on the rationale with which the *Cochrane collaboration* justifies the assessment of all research evidence available, including the so-called grey literature consisting of unpublished studies (Howick, 2011, p.15) [73]. So, to the extent that the members of the Cochrane collaboration endorse the PTE, their concern with systematic reviews has little to do with supporting clinical recommendations for individuals in accordance with the DA. After all, EBM's teachings and implementation

[73] Notice that although it is true that the Cochrane collaboration encourages reviewers to consider unpublished evidence, there are other reasons to think that neither the Cochrane collaboration nor EBM practice is truly committed to the PTE. I say this because it is well-known that the conduction of many systematic reviews involve not only the application of exclusion criteria to leave out studies that are irrelevant to the research question, but also the exclusion of relevant evidence that is discarded on the grounds that it is methodologically weak according to EBM standards. The application of these latter criteria, which cannot be defended in terms of irrelevance to the research question, cast doubt on the idea that the kind of systematic reviews supported by EBM are truly guided by the PTE.

strategies are committed to hierarchical rules of EBM that promote attention to selected research evidence rather than to all available information about the patient.[74]

So, to summarize, the claim that the DA lacks normative novelty because the EBM approach has already endorsed the PTE seems no more than a "red herring", for EBM does not truly support inferences based on this principle (See Ch.2. § 2.4).

Now, turning to the role assigned to clinical judgment, the fact that some supporters of EBM have mentioned clinical judgment as part of the EBM approach might lead some readers to think that DA's emphasis on clinical judgment again lacks originality, as it is only a repetition of an idea already accepted by the EBM approach. After all, a quick reading of EBM's standard definition (§ 2.3), which represents it as a judicious endeavour, might suggest that the EBM discourse is well aware of the importance of clinical judgment. But again, as previously argued (§ 2.4), although EBM's official definition seems to attribute clinical judgment a certain role, the 4S model makes it clear that the EBM version of judgment reduces its function to "applicability assessments", which are strictly subordinate to EBM rules of evidence.

It is worth emphasising that the role assigned to clinical judgment within the EBM camp is more heterogeneous than it first appears. In the opinion of many supporters of EBM, clinical judgment is above all (if not exclusively) concerned with the assessments of patients' preferences and values (e.g. Karthikeyan and Pais, 2010). Thus consider the advice of the most extreme branch of EBM, represented by authors such as Jeremy Howick (2011) and David Eddy (e.g. 2005): they have argued that clinical judgment should not be assigned any evidential role during clinical care. So, according to these authors, clinical judgment is highly misleading not only when used to back up general efficacy claims (as happens when expert judgment is used to justify therapeutic recommendations for populations), but also when backing up efficacy claims restricted to particular patients.[75]

So, in sum, the objection that the DA lacks normative novelty fails. Although ideas such as attending to all available evidence and exercising clinical judgment has been previously mentioned by some supporters of the EBM movement, the DA is more specific in its

[74] Furthermore, even if one could simply ignore the tension generated by hierarchical rules of evidence, I argued earlier that there are powerful incentives, which encourage physicians to restrict their attention to the information backed up by valid research (or simply to comply with EBGs) (Ch2. § 2.4). Thus, it is very difficult to see how the resulting clinical recommendations could be classified as emerging in accordance to the PTE.

[75] Both Jeremy Howick (2011) and David Eddy (1990a) contend that their claims are supported by empirical evidence. However, in the last sections of this chapter (§ 4.4) I shall question the quality and relevance of such evidence and reveal the need for new empirical evidence, which will be provided in last chapters of this thesis (5 and 6).

recommendations: first, the DA tells the physicians that the right probabilities ought to be based on everything the physician knows about the patient; and second, the DA is very explicit in that the role of clinical judgment is to estimate right probabilities for each patient.

4.3.2. Is it not what the average physician is already doing?

Let me now consider the objection that the DA only articulates more explicitly what the average physicians actually does in practice. According to this complaint, the DA lacks originality because its remarks on the right probabilities and the role of judgment only repeat elements presents in physicians' normal clinical practice.

This objection seems to find some support in the literature: various physicians have reported that they practice a type of medicine that is informed by research but fundamentally driven by their own judgment and clinical experience (e.g. Elwyn et al., 2016; McCartnery et al., 2016; Gupta et al., 2004; Grahame-Smith, 1995; Lancet editors, 1995; Tonelli, 1998). Nonetheless, the aforementioned references describe the practices of particular physicians, and this scarcely shows that all physicians proceed in this way. After all, it seems highly unlikely, given the widespread influence of the EBM movement[76], that most physicians would practice along the lines of the DA and not following the principles of EBM described in chapter one and two. Moreover, given that the system of EBGs is highly influential in countries with centralised healthcare systems such as the United Kingdom[77], with the resulting pressures attached to physicians' compliance with EBGs, it seems probable that most physicians follow according EBM's actual models of practice.

Of course, as we will see in the last two chapters of this dissertation, the existence of various initiatives to promote the implementations and compliance with EBGs is itself some evidence that there are still few physicians who may follow their own approaches to practice (e.g. personal guidelines), either in therapy (Chapter 5) or in the context of diagnosis and prognosis (Chapter 6). However, not practicing according to EBM is not necessarily synonymous with practicing along the lines of the DA. There are many ways of not complying with EBGs, and following the DA is only one of them (others include out-dated beliefs, laziness, carelessness, or any other reason different from basing recommendations in the right probabilities).

As far as I know, there is neither evidence nor principled reason to support the thesis that the DA truly represents how physicians practice, and to that extent the lack of descriptive novelty of this approach remains mere speculation. But notice that even if such evidence existed, I

[76] See Chapter 1 (footnote 3).

[77] I admit that the thesis that most physicians practice some sort of judgment-based medicine may hold true in some developing countries (Baradaran-Seyed et al., 2013) or in the context of private medicine (Lazarus, 2005).

would not be moved by it. From my perspective, a charge against descriptive novelty is simply misguided, for the importance and potential positive impact of the DA does not lie in its descriptive accuracy but rather on its normative influence: If proven reasonable and feasible, the DA might be able to tell physicians that the way in which EBM has taught them to practice medicine could be improved, and to the extent that such improvement is achievable, then the DA would be vindicated.

4.3.3. For a thoughtful physician, the DA is no more than a truism

Some practically-minded physicians might object that the DA is scarcely news. After all, if the only thing the DA tells physicians is that they ought to base their recommendations on everything they know about the patient and apply their judgment when the PEI arises, then, the DA would seem little more than a truism, from the perspective of any sensible physician.

My basic response to this objection is that, if the DA were indeed a truism for the sensible physician, I would be honoured rather than offended. What is more, I would be happy to know that there are physicians who do not require much thinking and argument to be convinced that clinical recommendations ought to be based on everything they know about their patients and that that their judgment plays a special role when comes to address the PEI.

Furthermore, another desirable consequence would be that, if it were true that the proposal behind the DA really were evidently unquestionable to the thoughtful physician, and in fact what the DA does is little more than encourage physicians to exercise their common-sense, then that would suggest that my main objection against the EBM could be easily fixed by helping sensible physicians –those who already endorse the DA– to instil a sufficiently significant dose of "common-sense" into the average physician's practice. And, of course, I would be very happy to accept such a scenario, even if the role of the DA is reduced to a mere reminder of what sensible physicians take as uncontroversial.

Regrettably, although I think that many thoughtful physicians would classify the DA as a matter of common sense, I believe that common sense is too vague an idea to describe precisely what the DA is advocating. True, it might be that common sense may hold the principles of the DA within it; that is, that common sense encompasses the DA. But still, to describe the DA in such terms would do little to fix the current situation of clinical medicine, which in my opinion is far more complex than just "a lack of common sense". So, even if the idea of basing recommendations on everything the physician knows about her patient in response to the PEI might seem trivial to the wise, it need not be trivial (and in fact it might seem controversial or even counterintuitive) to the common physician indoctrinated as an EBM practitioner. In fact, what I would expect from the average physician trained in EBM

would be very different sorts of criticisms against the DA, some of which I shall now turn my attention to: first, the DA is an approach that, in practice, cannot be executed by real physicians, and second, that, even if the DA were applicable in practice, physicians ought not to be practice it for it might worsen clinical care.

4.3.4. A good but impractical idea

The next objection is that, by stipulating that the right probabilities for clinical recommendations as those in the reference class defined by everything the physician knows about the patient, the DA is establishing too high a standard for the average physician. Even if the DA probabilities were the right probabilities to base clinical recommendations on, these recommendations would be, in practice, unattainable by real physicians in actual settings.

Perhaps, in an ideal world, where physicians posses unbounded attention and memory, extremely keen observation skills, as well unlimited time, the claim that DA probabilities are the right probabilities for clinical recommendations would have some practical import. However, so the objection goes, in the real world the DA advice lacks practical application.

I suspect, however, that this objection is based on a misunderstanding about what I am actually advocating for as a solution to the PEI. DA probabilities are the right probabilities from a prudential point of view, that is, physicians ought to be interested in them regardless of whether they are in practice capable of knowing about them. The value of this advice can only be understood in contrast to the mainstream idea that the right probabilities are what I have denoted EBM probabilities (Ch1, §1.5.1), which ignore part of the information available. Up to this point, what is at stake is what are the right probabilities, not whether physicians are actually capable of estimating them.[78] Nonetheless, even if epistemological limitations are a problem, it remains a valuable point to clarify that the right probabilities are *DA probabilities* rather than *EBM probabilities*. So, I think that to object that the notion of the right probabilities has no practical utility is to misunderstand its role, for knowing what are the probabilities of interest is of practical value regardless of the physicians' ability to estimate their value.

The real practical value in identifying the probabilities of interest, is to provide physicians with a *target* for the exercise of their judgment. According to the DA, the role of judgment is to address the PEI by estimating as best as possible the right probabilities for each individual. After all, sufficiently good estimate of the right probabilities is better than a perfect estimate

[78] This is a further matter, which has its own implications related to the various epistemological limitations that affect physicians during the clinical encounter.

of the wrong probabilities, given that the former are what physicians needs to arrive to the right recommendation for individual patients.

In this regard, let me stress that I acknowledge that physicians' judgment is subject to various practical constraints, which might make it difficult to estimate the right probabilities. However I do not think that the fact that judgment is affected by various limitations necessarily makes its exercise inadvisable. This is because this task is qualified in two important respects.

First, according to the DA, the probabilities of interest are those in the reference class defined by everything the physician knows about the patient. These probabilities are certainly very refined, but it must be stressed that these are not single probabilities, whose estimation is, without doubt, much more evidently unattainable to physicians because of the potentially endless number of factors unknown to the physician that may affect the outcome of interest in standard circumstances.

Second, while the DA prompts the obvious question of whether the physician should gather even more information before making a decision, and therefore threatens to turn clinical decision-making into a potentially endless gathering of further information, I have explained that rational physicians can address this issue by appealing to the Ramsey-Good result, which tells them that they ought to gather more information and defer recommendations only insofar the cost of doing so are worth it (footnote 22). This provides physicians with an intuitive answer to this challenge, which happens to be consistent with the observation that sensible physicians tend to decide when to gather more information by pondering the costs (broadly understood) involved in such action on a case-by-case basis.

Furthermore, with regard to the *estimation of the right probabilities*, it is worth stressing that the main point behind the DA, and in fact the point that captures what I think is the key weakness of the EBM approach, is that the presence of the PEI makes it unavoidable for the physician to take a view about the extra features present in the patient. But at the same time, this does not mean that the DA expects physicians to be able to come up with perfect numerical estimates of the right probabilities. As I said earlier, the DA is premised on the idea that an imperfect estimate of the right probabilities is generally better than a perfect estimate of the wrong probabilities. So, to repeat, a useful way to bear in mind the role assigned to judgment according to the DA is to recall that the solution to the PEI does not lie in exercising judgment to provide a perfect estimate, but in exercising judgment to estimate the right probabilities.

Of course, I agree with supporters of EBM that the DA remains largely silent as to how exactly judgment should be exercised, apart from the general advice that the physician has to make a judgment about the extra features she knows about the patient. I admit that this is not a precise recipe, but I do not think that the DA needs to provide physicians with a precise recipe in order to be practically useful. As I pointed out at the end of the first chapter, one of the virtues of accepting DA probabilities as the right probabilities for clinical recommendation is evidential flexibility (Ch1. § 1.5.2). This implies that physicians not only know that they have to exercise their judgment, but also that their judgment is not subservient to the rules of EBM.

True, this increased flexibility may be particularly challenging for inexperienced physicians, which might find specially difficult to make sense of extra information due to their lack of background knowledge. Furthermore, an additional challenge arises from the fact that there is no clear set of instructions for integrating different sources of information. Each of the extra factors present in the patient might have different weight, which could be very difficult to specify in particular if the factors are not probabilistically independent. For this reason, it is difficult to say something more specific than that the physician ought to start with certain prior probability for a set of hypotheses, and then that she ought to update the relevant probabilities on learning various facts about the patient so to address the PEI.

Nevertheless, even if it is unclear how physicians have to estimate the right probabilities, from my perspective it remains of practical utility to remind them that they will not do a favour to the patient if they, as the EBM approach advises, put their judgment to one side and just assume that the extra features present in the patient are probabilistically irrelevant.

At this point, however, supporters of EBM might raise a further (though related) objection: they might complain that if the DA is an approach that aspires to have some practical import on clinical care, we cannot simply put estimation problems to one side. After all, what is ultimately at stake is which of two practices will lead to the best clinical care: (a) exercising judgment so to address the PEI and estimate (as best as possible) the right probabilities or (b) ignoring the PEI and putting judgment to one side.

In the final section of this chapter I shall dissect what I call the “bad consequences” objection to the DA, which attacks the DA and supports the EBM approach by arguing that historical comparisons between physicians’ judgment and prediction methods purely based on statistical methods demonstrate the empirically bad effects arising from the limitations of human judgment.

4.4. The DA approach might worsen clinical care

An important source of concern about the DA is that its application might worsen the quality of clinical care. I this section I shall examine two claims underlying this worry. The first is that the DA is too permissive with regard to the epistemological import of clinical experience, and the second is that the DA is too optimistic about physicians' capacities to estimate probabilities.

Although I shall acknowledge that the DA might increase physicians' reliance on clinical experience, and that it would be unreal to expect that physician will base their clinical recommendations on perfect estimates of the probabilities of interest, I shall contend that neither of these consequences imply that the DA will worsen clinical care.

Nevertheless, I accept the aforementioned concerns as plausible worries, which need to be addressed with relevant empirical research comparing the effects of the DA and the EBM approach to the care of individuals.

4.4.1. The DA it is too permissive with respect to clinical experience.

In the first chapter, I pointed out that one of the positive by-products of directing physicians' attention to DA probabilities is "evidential flexibility" (§ 1.5.2). Given this feature, the DA allows physicians to incorporate information from clinical experience when addressing the PEI, which can be particularly useful to improve their estimates of the right probabilities. However, not everybody thinks of this feature of the DA as a virtue.

Supporters of EBM might be particularly reluctant to accept the DA's position on clinical experience for, in their opinion, it exposes the DA as a *laissez-faire* approach to clinical care that will allow physicians to break sensible rules of evidence. This is mistaken because it permits an excessive reliance on misleading evidence and simultaneously distracts physicians from the best guide available to clinical decision-making: valid research findings.

Consider the following remarks from David Eddy (1990a) a prominent supporter of EBM [79]:

> *"[Personal experiences] are notoriously misleading: the numbers of observations are small, there are no controls, patients and physicians decisions about interventions are not random, follow-up is incomplete and usually short term, and memories are highly selective."* (p.289).

Or the comments of Robyn Dawes and colleagues (1989):

[79] A useful summary of David Eddy's view on EBM and the version of practice he endorses (Evidence-Based Guidelines), can be found in Eddy 2005, 2011.

> *"The clinician is also exposed to a skewed sample of humanity and, short of exposure to truly representative samples, it may be difficult, if not impossible, to determine relations among variables."* (p.1671)

As these quotes show, hard-line supporters of EBM think that clinical experience is quite simply not to be relied upon, and therefore, clinical recommendations supported by it are likely to be the wrong clinical recommendations.[80]

Now, as I stressed in the first chapter when discussing "evidential flexibility", I do not deny that clinical experience can lead physicians astray (§ 1.5.2.3). As various authors have noticed (e.g. Dawes et al. 1989), clinical experience is typically based on unrepresentative samples, and because of this, I accept that it normally lacks generalizability. Furthermore, it has been known for a long time (e.g. Knapp et al. 1972) that standard clinical experience exposes physicians to "cases" not "controls", and so makes it very difficult for physicians to ascertain whether the signs, symptoms, or outcomes of interest occur more often among patients with or without the ailment.

However, I think that the limitations of clinical experience do not affect inferences which concern specific recommendations for particular patients, as much as inferences that are expected to support general, population-level, efficacy claims. Let me explain why.

First, in regard to the lack of representativeness of standard clinical experience, it is worth noting that the representativeness of a sample depends on what population one wants to make inferences about. So, although I take it as uncontroversial that standard personal clinical experience is a poor evidential source to back up general medical knowledge, the particular experience of a physician may well be sufficiently representative to back up specific recommendations deliberately restricted to the kind of patients the physician is used to taking care of (§ 1.5.2.3). For example, even if the diabetic patients normally seen by a physician are not random members of the population of diabetic patients in the UK, the selective nature of this physician's sample does not necessarily threaten the validity of the information she can obtain from her sample if the physician bears in mind that such information is in most cases only relevant to the kind of diabetics she generally sees, and is not necessarily transferable to diabetic patients from other localities (let alone to the population of diabetic patients in the UK). So, clinical experience can provide physicians with local knowledge about the particular spectrum of patients they serve (e.g. Poretsky 1985), and in the extreme case about particular

[80] For example, Jeremy Howick (2011) stresses the perils of attending to clinical experience by bringing attention to historical anecdotes such as the accidental introduction of the *"antenatal use of corticosteroids"*, which was initially resisted by many expert physicians, who allegedly did not trust the result of clinical trials in part because their own clinical experience did not support the idea that lung disease was a really a problem (p.161).

knowledge about very specific classes of individuals. Of course, such information is often not generalizable, but it remains useful, in particular because it is the kind of information (local patient idiosyncrasies) clinical trials would remain silent about.

Second, although the lack of controls that characterises clinical experience complicates the estimation of the evidential weight of potential interaction effects, in the context of medical prediction the physician might still use her personal experience to estimate the frequency of certain symptoms. If the physician is interested in the objective probability of the outcome of interest for the next diabetic patient coming to her clinic rather than for a diabetic patient randomly taken from the population, she may do well to rely on her clinical experience.

Relevant here are reports of concrete cases that support the claim that incorporating information from clinical experience can improve clinical care. For example, Crandall and Getchew-Reiter (1993) studied the detection of life-threatening infections in a neonatal intensive care unit (NICU). After observing that clinicians could detect infants developing such infections even before blood tests came back positive, the researchers interviewed clinicians and identified a range of indicators (cues) and sign combinations (patterns) not previously described in the literature. This supports the idea that clinical experience can provide useful information which can be incorporated in routine practice.

Likewise, a group of researchers studied 30 trauma cases and recorded the adherence of experts and novices respectively to the Advance Trauma Life Support standard (ATLS) (Kahol et al. 2011; Vankipuram et al. 2012). This study found that seasoned physicians' deviations from protocol were in most cases dynamic adjustments to adapt the standard guidelines and thus an improvement in efficiency and accuracy (Kahol et al. 2011; Vankipuram et al. 2012). Again, this supports the idea that physicians' attention to their clinical experience allows for flexibility and adaptiveness, which in turn may bring about better outcomes for patients.

So, although the aforementioned examples do not change the fact that clinical experience is limited in many respects and can be misleading, they do support the claim that sometimes it can be useful. Furthermore, it should not be forgotten that, when it comes to assessing the relative merits of clinical experience over other evidential sources, the kind of research evidence ranked at the top of hierarchies of evidence can be misleading too. It is well known that clinical trials, even those carefully conducted, are not free from errors related to statistical, causal inference or external validity, which introduces uncertainty as to their potential to guide practice. And, even more importantly, as I argued in the first chapter, even if there were no such problems and the physician had access to a perfect estimate, if such estimate ignores

part of what the physician knows about the patient, it would be an estimate of the "wrong probabilities" and may well lead to the erroneous recommendations.

So, to sum up, although I acknowledge that the DA might increase physicians' reliance on their own clinical experience, and that such experience has its limitations, I do not think that physicians should automatically jettison such kind of information. This would not only be prudentially irrational but also, it might hamper physicians' capacity to deal effectively with the PEI and therefore decrease their capacity to estimate the right probabilities.

4.4.2. The DA underrates physicians' limited capacities to estimate probabilities

The DA entrusts physicians with an important responsibility: whenever they know something extra about the patient, they ought to exercise their judgment to estimate the right probabilities. But what if physicians are such poor estimators that, even if the DA directs their attention to the right probabilities the exercise of their judgment is incapable of arriving a minimally appropriate estimate?

Backing for this worry is provided by the heuristics and biases tradition of research in psychology (Tversky and Kahneman, 1974; Tversky, 2004; Samuels et al., 2002; Papineau, 2006). This research is extensive, controversial and with wide implications on many different topics. For the purposes of this chapter I shall limit myself to (i) admitting that there are various biases affecting human judgment (some of which are specifically relevant to our discussion), (ii) contending that the impact of these biases on physicians' clinical decision making is far from clear, and therefore that (iii) it is reasonable to think that supporters of EBM (e.g. Howick, 2011) have overstated the perils of physicians' judgment in the context of clinical medicine.[81]

Let me start by describing, without attempting to be exhaustive, a small list of relevant biases that are relevant to the PEI. For a start, various authors (e.g. Elstein 1990, 1999) have provided evidence that (i) humans tend to over-attend to information consistent with their own hypotheses and to under-attend to contradictory information.[82] Likewise, it has been documented that physicians are normally victims of overconfidence: that is, they typically overestimate the accuracy of their judgments (e.g. Berner and Graber, 2008; Elstein, 1999). In addition, it is now almost common knowledge that many physicians disregard frequency data

[81] Notice that a potential line of argument to defend the DA against the objection that assigning such an important role to clinical judgment is erroneous would be to attack the external validity of the experiments provided by the psychologists in the heuristics and biases tradition in general. I shall not pursue this line argument because I shall focus on more specific objections.

[82] As Grove and Meehl (1996) says "this is simply the ineradicable tendency of the human mind to select instances for generalizations that it favours" (p.15).

and often makes predictions on the basis of some prototype or instance stored in memory (Kahneman and Tversky, 1973; Pennycook et al. 2014; Charlin et al., 2007).

However, although the aforementioned criticisms obviously raise doubts about the reliability of human judgment, it is unclear whether they necessarily lead to significant practical problems in the context of clinical care. For example, take the claim that physician tend to focus on part of the information available. While there is a sense in which this raises doubts about the capacity of physicians to successfully address the PEI, its practical consequences remain unclear. Studies on diagnostic reasoning suggest that it is unclear whether successful diagnosticians do this less than unsuccessful ones (Eva, 2005; Norman, 2005, Norman and Eva 2003; Ericsson, 2007). In fact, when we do this and the outcome is good we typically call it heuristic, but if we do the same but the outcome is bad, we then call it premature closure bias. Both labels, however, are instances of wisdom after the event, and there is no strong empirical evidence suggesting that partial focus is necessarily bad in natural settings (Norman, 2014).

Physicians' overconfidence, on the other hand, does sound as if it will be problematic. However, the real issue is whether such overconfidence really does lead to worst clinical recommendations. Sometimes it is precisely overconfidence that allows the physician to exercise her judgment and focus on the most relevant elements of the case. Such a procedure is neither inconsistent with the DA, nor necessarily bad, for it is by no means clear whether it leads to systematic error (See Croskerry and Norman, 2008).

Automatic disregard of base-rate fallacies is considered a serious error in probabilistic reasoning. However, one must be careful not to confuse this kind of bias with situations, which may be common in real-life, where the physician prefer to ignore base-rates from the literature, in favour of more idiosyncratic or local base-rates, which may often be more relevant to calculating the right probabilities for the patient. For example, a clinician who works in a penitentiary or forensic psychiatric clinic might conclude, after considered judgment, that the frequency of liars among psychiatric patients is very low. This is not a case of base-rate fallacy, but a case where the physician judges such information to be irrelevant to the probability for the patient she has in front of her—the fact that the patient has been hospitalised in a forensic psychiatric service changes the relevant base rate.

So, although I do not deny that physicians' judgment is vulnerable to various sorts of bias, I do not think it reasonable to conclude that the right attitude towards that problem is to abandon judgment altogether, and automatically apply valid research data on the basis of common membership to general diagnostic reference classes. After all, leading authors in the heuristic and biases tradition, and also in the field of expert knowledge, concur on the idea

that these biases do not necessarily prevent arriving at the right recommendations (Kanehman and Klein, 2009).

More fundamentally, if the main question is whether physicians' biases are so severe as to justify the abandonment of judgment and support clinical recommendations based exclusively on valid research data, as the extreme branch of EBM proposes (Howick, 2011), then documenting biases is not enough to answer this question, for what is really needed is comparative empirical evidence that predictions and prescriptions based on research data and without the interference of judgment are superior to those where clinical judgment is used to address the PEI.

In this respect, some supporters of EBM (Howick, 2011) think that such evidence is already available and that it is sufficient to justify the otherwise controversial claim that clinical judgment "*belongs to the bottom of the hierarchy [of evidence]"* (p.167).

In the last section of this chapter, I shall comment on the so-called actuarial versus clinical prediction research tradition, launched by the prominent psychologist Paul Meehl during 1950s, and which is used by Jeremy Howick (2011) to substantiate his claims that, even if there are problems with the applicability of research, there is no reason to think that increasing the room for clinical discretion, as the DA recommends, would improve clinical care.

4.4.3. The need for relevant and updated empirical comparisons

When defending claims such as *"where high-quality comparative clinical trials do exist, data should trump the judgment of experts in diagnostic, prognostic and therapeutic predictions"* (Howick, 2011, p.166), or *"allowing clinicians to "break" the mechanical rule will tend to make outcomes worse"* (Howick, 2011, p.169) or that judgment ought to be restricted to *"non-evidential roles"* [83], supporters of the more extreme branch of EBM assert that the EBM position *"is well supported"* (Howick, 2011, p.161).

[83] By a *"non-evidential role"* in the context of decision-making for individuals, Jeremy Howick (2011) literally means that *"…expert judgment should not be used as evidence"* (p.177) and that the function of the physician is to integrate *"the best research evidence with patient values and circumstances"*(p.177). Notice that the word *"circumstances"* in Howick's usage should not be confused with my use of the word *"information"*, such as when I say that the objective probabilities of interest for the physicians are those in the reference class defined by all *"information"* about the patient available to him. While I interpret "*information"* in an evidential sense, that is, as an evidential input that aids the physician in the estimation of the right probabilities, Jeremy Howick speaks about the "*patient's values and circumstances"* to convey the idea that the role of judgment is to consider what kind of interventions are both consistent with the *patient's interests* and *practically possible* or *feasible.* Obviously the latter is something independent of the probability that the intervention in question will cause the outcome of interest for the patient).

According to Jeremy Howick (2011): *"The hypothesis that expert judgment is useful in applying average statistical results to individual patients has been tested hundreds of times over the course of the last six decades. [But] with a mere handful of exceptions, they all suggest that following "mechanical" rules is as good as, or better than, expert judgment…"* (p.167).

In this final section, I am going to argue against Howick's claims. My aim is to show that his assessment and interpretation of the evidence he cites in support of the EBM position on clinical judgment is deficient and that most of his analysis is unsound. However, my conclusion will not be that we should take the virtues of clinical judgment for granted, but rather that such virtues, and the virtues of an alternative approach such as EBM, deserve to be examined via updated and methodologically relevant empirical evidence.[84]

The literature Howick cites to support the lack of utility of physicians' judgment in clinical settings belongs to the "*Actuarial versus Clinical Prediction*" [85] tradition, launched by the psychologist Paul Meehl in his book *"Clinical versus Actuarial Prediction a theoretical analysis and a review of the evidence"* (Meehl, 1954).

In his discussion of the experiments reported in Meehl's book (1954), Howick (2011) ignores the fact that for Meehl the question of interest was not whether clinicians [86], given their various cognitive limitations, would be able to address the PEI and thereby provide patients with better predictions and prescriptions. Rather Meehl was interested in a different question, namely: whether clinicians or mechanical rules were better at combining the same data. In Meehl's words (1954), he was interested in investigating *"the relative efficiency of actuarial and non-actuarial methods of combining the same data to yield a prediction"* (p.118).

Meehl's question is obviously a valid one, and particularly understandable once one considers the historical context in which he carried out his research was one where the prevailing assumption was that expert judgment, rather than research evidence, constituted the best source of general (population-level) knowledge at least in psychology and medicine. [87]

[84] To be fair, in his most tempered moments Jeremy Howick (2011) acknowledges at least some of the limitations of the evidence he cites, but for the most part he fails to see that his contentions are not well supported by the empirical data he offers.

[85] The term "actuarial" was quickly replaced by "mechanical" and later on by "statistical". So the current debate is between the merits of "statistical" and "clinical" prediction.

[86] The term "clinicians" is a broader term used to describe workers of different backgrounds and educational levels who perform clinical work, usually physicians, psychologists, and nurses but also sometimes biochemists, medical technicians, and even sociologists who have direct contact with patients.

[87] One should not forget that Paul Meehl was an experimental psychologist, which was educated in a time when psychoanalytical theory was taught and applied in the United States as if it were almost an unquestionable truth. In that context, the psychoanalytic expert himself, represented the most sophisticated source of general knowledge,

However, as I said, Meehl's question is very different to the question of whether addressing the PEI by increasing the room for clinical discretion results in better predictions and prescriptions. In fact, the kind of experiments that address Meehl's research question artificially eliminate the PEI, by considering how the same data is processed by statistical rules and physicians.

Of course, this observation does not invalidate all experiments cited by Howick (2011). But we should be suspicious of his assertion that *"the EBM position on expert judgment as evidence is well supported by a plethora of largely ignored studies…"*(p.161). Much of Howick's plethora of studies is simply irrelevant to the point under discussion.

As it happens, Howick also cites the research conducted by William Grove, a disciple of Paul Meehl (Grove et al. 2000). This was designed to examine the relative predictive performance of physicians' judgment and statistical models from a broader perspective, and is indeed relevant to EBM's general position on clinical judgment. However, Howick overlooks some crucial problems that limit both the validity and relevance of Grove's findings to clinical medicine. Since these limitations are not restricted to Grove's meta-analyses and also affect the work of other researchers in this area (Dawes et al, 1989; Marchesse, 1992; Ægisdóttir et al. 2006), it is worth explaining why they render much of this research irrelevant to the assessment of EBM.

For a start, most of the studies cited were conducted using case summaries instead of real cases. This is not trivial for several reasons. In the first place, case summaries artificially eliminate the PEI. In addition, the use of case summaries limits the external validity of the findings: as is well known, physicians rely on many sources of information, and not all of them are easily described let alone included in a case-summary (Patel and Groen, 1991). Moreover, to examine the merits of statistical predictions (SP) and clinical predictions by physicians (CP) on the basis of restricted datasets is also inadequate because another argument raised against judgment is that physicians' predictive performance might be affected by *"dilution effects"* (Nisbett et al. 1981; Zukier, 1982; Tetlock and Bottger, 1989), which are also missed by experiments using case summaries. This is particularly important for the prospects of the DA because, if it is true that physicians facing real patients tend to pay attention to false clues and ignore relevant ones, then the accuracy of physicians' estimates of the right probabilities would be highly diminished. So, given that, as Rakow and colleagues (2005) say: *"[the] asymmetry in the information available to the doctors and the*

---

that is, the best evidential source someone can hope for. The reader interested in these points is invited to read the first sections of Meehl's book (1954, pp. 3 to 82).

prediction models merely reflects the inherent constraints and advantages in the everyday use of each approach" (p. 262), there are reasons to question the validity of studies based on case summaries, and pay particular attention to comparisons conducted in real settings, where physicians really confront the PEI, and their estimation skills are truly put to the test. [88] Of course, the practical success of the DA with respect to the estimation of the right probabilities assumes inter alia that when the physician is confronted with an actual patient, advantages such as the access to the patient's characteristics will typically counterbalance disadvantages such as the physician's cognitive overload, or potential "dilution effects", but this hypothesis needs to be examined, and that cannot be done in artificial settings with case summaries.

Another set of problem with previous studies comparing SP with CP is that in many of them (a) the outcome of interest did not pertain to clinical medicine, (b) the predictive tasks were outside the medical domain or (c) the tasks did not involve physicians. As interesting as it may be to know that certain algorithms predict better than humans outcomes such as "voting behaviour", "job satisfaction", "next day's weather", "marriage success", "adjustment to prison life", or "parole violations" (Grove et al. 2000), the superiority of SP over CP reported by primary studies focused on the aforementioned outcomes is not relevant to the question of whether the predictive performance of physicians' judgment is superior or inferior to that of statistical models. Furthermore, since the scope of application of both the EBM approach and physicians' judgment does not extend to non-clinical settings, primary studies focused on *military personnel* or *graduate students* are scarcely useful to either substantiate the claims defended by EBM or support my defence of physicians' judgment.

Likewise, since there is some evidence suggesting that the judgment of members of different professions has different predictive abilities (Kaufmann and Athanasou 2009, Kaufmann et al., 2013), and that the level of experience might affect predictive performance (LaDuca et al., 1988, Smith et al., 2003), studies that report that the predictive performance of *business managers*, *social workers*, *psychoanalytic psychotherapists*, *members of parole boards*, and *medical students* is inferior to that of statistical models is loosely connected to the claim on which the practical success of either the EBM approach or the DA is premised. To repeat, what we are interested in is in primary studies showing that when properly qualified physicians (not medical students) apply their judgment they do not do it in such way that results in a greater number of erroneous predictions than the corresponding number of erroneous predictions that follow from the strict application of statistical algorithms. So, my

[88] Notice a special case of extra information that might help physicians, which is not captured by case summaries, is the information about local patters of disease. As we will see in the last chapter of this thesis (Chapter 6), deeper knowledge about the kind of patients that normally visit their clinic seems to be an important comparative advantage for physicians.

point is simply that if one wants to investigate the merits of judgment to handle PEI, we need to focus on relevant studies (that is, studies that include physicians predicting clinical outcomes).

Finally, another, more technical, limitation that weakens the validity of several studies allegedly supporting the EBM's position on judgment, and one that was ignored by Jeremy Howick's analysis, was that in many primary studies the performance of SP was measured in derivation samples. As several authors have pointed out, one should not assess the performance of a statistical model on the patients on which the model was first created, because it is to be expected that the model will perform better on the data from which it was derived than on any new sample, regardless of the fact that new samples have been drawn from the same underlying population (e.g. Dawid 1976; Spiegelhalter and Knill-Jones 1984; Bleeker et al., 2003). This is because the derivation of statistical rules might be based on non-repeating relations among variables, and as a result, the estimates obtained by many researchers are often affected by problems such as "over-fitting".[89] So, the point here is that several studies cited by Howick were studies in which the performance of the models was measured using estimates obtained during the derivation process and therefore overestimate the performance of such models.[90]

Thus, although the aforementioned problems do not entirely invalidate Howick's defence of the EBM position on clinical judgment, at the very least they cast doubt on the soundness of his analysis.

So, to sum up, careful attention to what Howick describes as *"a plethora of largely ignored studies"* supporting the EBM's position on clinical judgment reveals that many of those studies are irrelevant, methodologically questionable or both. Of course, some of the evidence cited by Howick might be relevant and informative in assessing the merits of physicians' judgment and the practical prospects of the DA. However, my analysis suggest that the data presented by Howick are far less compelling than he suggests, and for this reason I do not think it provides supporters of EBM with a strong case against physicians' judgment, let alone against the DA.

[89] For a clear exposition of the problem of "overfitting", which from the philosophical point of view is captured by the distinction between predicting and accommodating see (Hitchcock and Sober 2004)

[90] Notice that "new sample" here is not used to describe a new sample composed by obviously different patients but rather a new sample composed by members of the original target population. In this respect several authors have reported actual cases where the predictive performance of statistical models applied to new samples of patients drops significantly (Toll et al. 2008; Justice et al. 1999).

Since the superiority of SP over CP, and more generally of the EBM approach over the DA is an empirical matter, the sensible next step is to conduct more careful and systematic investigation of the practical performance of both EBM and the DA. In the end, what is at issue here is whether EBM –an approach that favours automatic reliance on statistical rules– is actually more beneficial to patients than the DA –an alternative approach which attends to the PEI and calls for clinical discretion.

4.5. Conclusions

The first aim of this chapter was to demarcate the DA from apparently similar positions, which nevertheless have rather different implications for clinical care. It was observed that the DA ought not to be confused with MP, a position which questions all inferences based on general classes, nor with PCT, an approach based a broad but indefinite notion of adequate care.

In addition, this chapter considered a set of concerns as to the practicalities involves in the DA. I argued that the precision of the DA ensured its normative novelty. In addition, I contended that a charge descriptive unoriginality is unjustified for it is plausible to think that the average physician nowadays practices clinical medicine according to the standard models of EBM. I then accepted that from the perspective of the thoughtful experienced physician the postulates of the DA might be truistic, and observed that this was scarcely an objection to the DA. As a first response to the charge of impracticality, I observed that a rough estimate of the right probabilities can well be better than a perfect estimate of the wrong ones.

Finally, the last sections of this chapter were concerned with the possibility the DA will in practice lead to an increased number of wrong decisions. I observed that there was no principled reason why the DA should have this consequence. I then added a critical appraisal of the literature cited by supporters of EBM to substantiate the claim that clinical judgement was generally inferior to statistical prediction. I showed that such literature has serious methodological limitations, and that what is needed is a methodologically sound comparison between the DA and the EBM approach. This will be the subject of the last two chapters of this thesis.

4.6. References

Baradaran-Seyed, Z., Nedjat, S., Yazdizadeh, B., Nedjat, S. and Majdzadeh, R. (2013). Barriers of clinical practice guidelines development and implementation in developing countries: a case study in iran. *Int J Prev Med.* 4(3):340-8.

Berner, E.S. and Graber, M.L. (2008). Overconfidence as a cause of diagnostic error in medicine. *Am J Med.* 121(5 Suppl):S2-23.

Bleeker, S.E., Moll, H.A., Steyerberg, E.W., Donders, A.R., Derksen-Lubsen, G., Grobbee, D.E. and Moons, K.G. (2003). External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol.* 56(9):826-32.

Bullock, E. and Kingma, E. (2014). Interdisciplinary workshop in the philosophy of medicine: medical knowledge, medical duties. *J Eval Clin Pract.* 20(6):994-1001.

Charlin, B., Boshuizen, H.P., Custers, E.J. and Feltovich, P.J. (2007). Scripts and clinical reasoning. *Med Educ.* 41(12):1178-84.

Crandall, B. and Getchell-Reiter, K. (1993). Critical decision method: a technique for eliciting concrete assessment indicators from the intuition of NICU nurses. *ANS Adv Nurs Sci.* 16(1):42-51.

Croskerry, P. and Norman, G. (2008). Overconfidence in clinical decision making. *Am J Med*. 121(5 Suppl):S24-9.

Dancy, J. (2004). *Ethics Without Principles*. Oxford, Oxford University Press.

Dawes, R.M., Faust, D. and Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science.* 243(4899):1668-74.

Dawid, A.P. (1976). Properties of diagnostic data distributions. *Biometrics.* 32(3):647-58.

Eddy, D.M. (1990). The Challenge. *JAMA.* 263(2):287-90.

Eddy, D.M. (2005). Evidence-based medicine: a unified approach. *Health Aff (Millwood).* 24(1):9-17.

Eddy, D.M. (2011). The origins of evidence-based medicine--a personal perspective. *Virtual Mentor.* 13(1):55-60.

Ægisdóttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook R.S., Nichols, C.N., Lampropoulos, G.K., Walker, B.S., Cohen, G. and Rush, J.D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist.* 34(3):341-82.

Elstein, A.S. (1990). Framing effects. *Med Decis Making.* 10(2):148.

Elstein, A.S. (1999). Heuristics and biases: selected errors in clinical reasoning. *Acad Med.* 74(7):791-4.

Elwyn, G., Wieringa, S. and Greenhalgh, T. (2016). Clinical encounters in the post-guidelines era. *BMJ.* 353:i3200.

Ericsson, K. A. (2007). An expert-performance perspective of research on medical expertise: The study of clinical performance. *Med Educ.* 41(12):1124-30.

Eva, K.W. (2005). What every teacher needs to know about clinical reasoning. *Med Educ*. 39(1):98-106. [Erratum, Med Educ 2005; 39:753.]

Eva, K. and Cunnington, J. (2006). The difficulty with experience: does practice increase susceptibility to premature closure? J *Contin Educ Health Prof.* 26(3):192-8.

Grahame-Smith, D. (1995). Evidence based medicine: Socratic dissent. *BMJ.* 310(6987):1126-7.

Grove, W.M. and Meehl, P.E. (1996). Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical–Statistical Controversy. *Psychology, Public Policy, and Law.* 2:293-323.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E. and Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychol Assess.* 12(1):19-30.

Gupta, M. (2004). Evidence-based medicine: ethically obligatory or ethically suspect? *Evid Based Ment Health.* 7(4):96-7.

Hitchcock, C. and Sober, E. (2004). Prediction Versus Accommodation and the Risk of Overfitting. Br J Philos Sci. 55(1):1-34.

Howitz, J. (2011). *The Philosophy of Evidence-based Medicine.* Oxford, Wiley-Blackwell.

ICPCM. (2016). "*Person-centered Medicine*". Retrieved 05 March, 2016, from: http://www.personcenteredmedicine.org/

Justice, A.C., Covinsky, K.E. and Berlin, J.A. (1999). Assessing the generalizability of prognostic information. *Ann Intern Med.* 130(6):515-24.

Kahol, K., Vankipuram, M., Patel, V.L. and Smith, M.L. (2011). Deviations from protocol in a complex trauma environment: errors or innovations? *J Biomed Inform.* 44(3):425-31.

Kahneman, D. and Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *Am Psychol*. 64(6):515-26.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological Review.* 80(4):237-51.

Kaltoft, M.K, Nielson, J.B., Eiring, O., Salkeld, G. and Dowie, J. (2011). Without a reconceptualisation of 'evidence base' evidence-based person-centred healthcare is an oxymoron. *European Journal for Person Centered Healthcare.* 3(4).

Karthikeyan, G. and Pais, P. (2010). Clinical judgement & evidence-based medicine: time for reconciliation. *Indian J Med Res.* 132:623-6.

Kaufmann, E. and Athanasou, J. (2009). A Meta-Analysis of Judgment Achievement as Defined by the Lens Model Equation. *Swiss Journal of Psychology*. 68(2):99-112.

Kaufmann, E., Reips, U.D. and Wittmann, W.W. (2013). A critical meta-analysis of lens model studies in human judgment and decision-making. *PLoS One.* 8(12):e83528.

Kitson, A., Marshall, A., Bassett, K. and Zeitz, K. (2013). What are the core elements of patient-centred care? A narrative review and synthesis of the literature from health policy, medicine and nursing. *J Adv Nurs.* 69(1):4-15.

Knapp, M.S., Cove-Smith, R., Hall, G. and McIllmurray, M. (1972). Dangers of diazoxide. *Br Med J.* 4(5834):229-30.

LaDuca, A., Engel, J.D. and Chovan, J.D. (1988). An Exploratory Study of Physicians' Clinical Judgment: An Application of Social Judgment Theory. *Eval Health Prof.* 11(2):178-200.

Lancet Editors (1995). Evidence-based medicine. *Lancet*. 346:785.

Lazarus, J.A. (2005). *Entering Private Practice: A Handbook for Psychiatrists.* Washington, American Psychiatric Publishing.

Loughlin, M. (2009). The search for substance: a quest for the identity- conditions of evidence based medicine. *J Eval Clin Pract.* 15(6):910-4.

Loughlin, M. (2014). What person-centred medicine is and isn't: temptations for the 'soul' of PCM. European Journal for Person Centered Healthcare. 2(1):16-21.

Marchese, M.C. (1992). Clinical versus actuarial prediction: a review of the literature. *Percept Mot Skills.* 75(2):583-94.

McCance, T., McCormack, B. and Dewing, J. (2011). An exploration of person-centredness in practice. Online J Issues Nurs. 16(2):1.

McCartney, M., Treadwell, J., Maskrey, N. and Lehman, R. (2016). Making evidence based medicine work for individual patients. *BMJ.* 353:i2452.

Meehl, P.E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, University of Minnesota.

Mezzich, J. E., Snaedal, J., van Weel, C. and Heath, I. (2010). *Toward person-centered medicine: from disease t*o patient to person. Mt Sinai J Med. 77(3):304-6.

Mezzich, J.E., Snaedal, J., van Weel, C., Botbol, M. and Salloum, I. (2011). Introduction to person-centred medicine: from concepts to practice. J Eval Clin Pract. 17(2):330-2. doi: 10.1111/j.1365-2753.2010.01606.x.

Miles, A. and Loughlin, M. (2011). Models in the balance: evidence-based medicine versus evidence-informed individualized care. *J Eval Clin Pract.* 17(4):531-6.

Miles, A. and Mezzich, J.E. (2011). The care of the patient and the soul of the clinic: person-centered medicine as an emergent model of modern clinical practice. *International Journal of Person Centered Medicine.* 1(2):207-22.

Miles, A. (2012). Person-centered medicine-at the intersection of science, ethics and humanism. *International Journal of Person Centered Medicine*. 2(3):329-33.

Norman, G. and Eva, K.W. (2003). Doggie diagnosis, diagnostic success and diagnostic reasoning strategies: an alternative view. *Med Educ*. 37(8):676-7.

Norman, G. (2005). Research in clinical reasoning: past history and current trends. *Med Educ.* 39(4):418-27.

Norman, G. and Eva, K. (2010). Diagnostic error and clinical reasoning. *Med Educ.* 44(1):94-100.

Norman, G. (2014). The Bias in researching cognitive bias. *Adv Health Sci Educ Theory Pract.* 19(3):291-5.

Nisbett, R.E., Zukier, H and Lemley, R.E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. Cognitive Psychology. 13(2):248-77.

Papineau, D. (2006). *The Roots of Reason: Philosophical Essays on Rationality, Evolution, and Probability.* New York, Oxford University Press.

Patel, V.L. and Groen, G.J. (1991). *The general and specific nature of medical expertise: A critical look.* In Toward a general theory of expertise: Prospects and limits. K.A. Ericsson and J. Smith (eds). New York, Cambridge University Press. pp. 93-125.

Pennycook, G., Trippas, D., Handley, S.J. and Thompson, V.A. (2014). Base rates: both neglected and intuitive. *J Exp Psychol Learn Mem Cogn*. 40(2):544-54.

Poretsky, L., Leibowitz, I.H. and Friedman, S.A. (1985). The diagnosis of myocardial infarction by computer-derived protocol in a municipal hospital. *Angiology*. 36(3):165-70.

Rakow, T., Vincent, C., Bull, K. and Harvey, N. (2005). Assessing the likelihood of an important clinical outcome: new insights from a comparison of clinical and actuarial judgment. *Med Decis Making.* 25(3):262-82.

Roberti di Sarsina, P. and Iseppato, I. (2010). Person-centred medicine: towards a definition. *Forsch Komplementmed.* 17(5):277-8.

Roberti di Sarsina, P., Alivia, M. and Guadagni, P. (2012). Traditional, complementary and alternative medical systems and their contribution to personalisation, prediction and prevention in medicine-person-centred medicine. *EPMA J.* 3(1):15.

Samuels, R., Stich, S. and Bishop, M. (2002). *Ending the Rationality Wars: How to Make Disputes About Human Rationality Disappear*. In *Common Sense, Reasoning and Rationality.* R. Elio (ed). New York: Oxford University Press. pp. 236-68.

Smith, G.C. and Pell, J.P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ*. 327(7429):1459-61.

Spiegelhalter, D.J. and Knill-Jones, R.P. (1984). Statistical and Knowledge-Based Approaches to Clinical Decision-Support Systems, with an Application in Gastroenterology. *Journal of the Royal Statistical Society. Series A (General)*. 147(1):35-77.

Tetlock, P.E. and Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology*. 57(3):388-98.

Toll, D.B., Janssen, K.J., Vergouwe, Y. and Moons, K.G. (2008). Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol.* 61(11):1085-94.

Tonelli, M.R. (1998). The philosophical limits of evidence-based medicine. *Acad Med.* 73(12):1234-40.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and Biases. *Science*. 185(4157):1124-31.

Tversky, A. (2004). *Preference, Belief, and Similarity: Selected Writings.* Massachusetts, The MIT Press.

Vankipuram, M., Ghaemmaghami, V. and Patel, V.L. (2012). Adaptive behaviors of experts in following standard protocol in trauma management: implications for developing flexible guidelines. *AMIA Annu Symp Proc*. 2012:1412-21.

Zukier, H. (1982). The dilution effect: The role of the correlation and the dispersion of predictor variables in the use of nondiagnostic information. *Journal of Personality and Social Psychology.* 43(6):1163-74.

Chapter 5: EBM guidelines vs. "usual care": A re-analysis of systematic reviews

5.1. Abstract

The Evidence-based Medicine (EBM) approach to medical practice encourages physicians to comply with Evidence-based Guidelines (EBGs) based on the best research evidence available (2.4.2). One of the fundamental assumptions of the EBGs model of care is that physicians' compliance with EBGs leads to better health outcomes. Throughout this dissertation I have cast doubt on this assumption by arguing that the correct recommendations for individuals need not be based on research probabilities. Instead I argued that research evidence ought to inform care but that sound clinical recommendations ought require the exercise of physician' judgment on the basis of everything she knows about the patient.

The main aim of this chapter is to compare the performance of successfully implemented EBGs with that of "care as usual" with respect to patient outcomes in the hope that this comparison will shed empirical light on the relative therapeutic virtues of the EBM approach and the DA. With this purpose in mind I shall conduct a reanalysis of systematic reviews investigating the efficacy of EBGs versus "care as usual" as proxies for the aforementioned approaches.

In this study, eight systematic reviews of the effectiveness of EBGs published from 1993 onwards were identified and analysed. These reviews reported the results of 149 primary studies, all of which were assessed in full. Forty-two primary studies measured *patient outcomes* (28.2%), and only ten of them met the methodological requirements for comparing EBGs with "care as usual" as a *proxy* of the DA. Among them, eight studies indicated that EBGs were either ineffective (n= 4) or mostly ineffective (n=4) with respect to patient outcomes. Only two studies favoured the application of EBGs over "care as usual".

This reanalysis of systematic reviews indicates that therapeutic recommendations based on EBGs does not necessarily improve patient outcomes. In fact, the data collected suggest that *"care as usual",* which relies on clinical discretion, usually obtains better patient outcomes than those obtained by EBGs. To the extent that the successfully implemented EBGs and "care as usual" are reasonable proxies for the EBM approach and the DA, this reanalysis challenges the superiority of the EBM approach, and answers the concern that adopting the DA would decrease in the quality of care.

5.2. Background

5.2.1. The rationale behind EBGs

Between 1980 and 1990, when David Sackett and colleagues (EBM group 1990), were developing the fundamental principles of the 4S model of EBM practice (§ 2.4.1), David Eddy, the physician who first used the term "evidence-based" in the medical literature, was aiming to persuade policy-makers, medical associations, and physicians themselves that clinical care should be standardised on the basis of Evidence-Based guidelines (EBGs) (Eddy 1982, 1990a-e, 2005, 2011).

Eddy's first task was to question the quality of clinical decisions (Eddy, 1990a). To do so, he directed attention to what later came to be called *"unwarranted variation"* in healthcare (Wennberg, 2002). This concept aimed to establish a link between a lack of an obvious explanation for many clinical recommendations and poor-quality clinical care. In Eddy's Words (1990):

"The plain fact is that many decisions made by physicians appear to be arbitrary–highly variable, with no obvious explanation. The very disturbing implication is that this arbitrariness represents for at least some patients, suboptimal or even harmful care" (p.289)

Eddy's understanding of variation in healthcare focused on the limitations of physicians' judgment. From his perspective, this was only capable of following simple *"if-then"* rules (Eddy, 1982), and–to make things worse–was informed by the wrong evidential sources (e.g. clinical experience and physiopathological rationale) (§ 1.5.2 and § 4.4). So, for instance, with regard to physicians' capacity to determine the right decisions, Eddy asserts (1990a):

"It is easy to appreciate that if a physician's perception of the outcomes of alternative interventions is incorrect, the chance that he or she will choose the best intervention for a patient is severely threatened." (p.287).[91]

Supporting Eddy's ideas, John Wennberg (2011)–another prominent figure in the area of EBGs and the researcher who provided much of the empirical evidence supporting the existence of *variation in care* (e.g. Wennberg, 1973, Wennberg et al. 1989)–assumed that *"Much of the variation in use of healthcare is accounted for by the willingness and ability of doctors to offer treatment rather than differences in illness or patient preference"* (p.687) and

[91] David Eddy backed up his assertion that physicians provided "inappropriate" care with studies suggesting that physicians were inconsistent in their recommendations for what appeared to be "similar" patients (e.g. Chassin, 1987).

that the majority of the recommendations made by physicians were biased, poorly justified, and likely to lead to bad patient outcomes (Wennberg, 2002).

Against this backdrop, the standardisation of care via EBGs emerged as a reasonable approach to improve the quality of care (Bodgan-Lovis et al. 2012, Tannenbaum, 2012; Eddy, 2011). For, from this perspective, with proper practice standardization on the basis of the best research evidence, most of the "unwarranted" differences in care could be eliminated (Claridge, 2005; Eddy, 2011; Kinney, 2004).[92]

5.2.2. A problematic assumption of the EBGs programme

As Avendis Donabedian (2005)—a renowned researcher in the field of quality of care—warned: "*the process of evaluation [of healthcare] itself requires much further study. A great deal of effort goes into the development of criteria and standards which are presumed to lend stability and uniformity to judgments of quality and yet this presumed effect has not been empirically demonstrated."* (p.715-716).

So, as Donabedian suggests, it is naïve to think that standardising healthcare interventions would immediately improve the quality of care. The situation is obviously much more complex, and therefore are problems with aiming to improve clinical care simply by *"getting research intro practice"* (e.g. Haines and Donald, 2002; Garner et al., 1998; Haynes and Haines, 1998). This assumes that the main failures related to improving the "science" backing up clinical recommendations (Sackett and Rosenberg, 1995) or overcoming *"implementations problems"* (Grimshaw, 1995). Again, Donobedian (2005) expressed his worries eloquently: *"One must also consider whether, with increasing standardization, so much loss of the ability to account for unforeseen elements in the clinical situation occurs that one obtains reliability at the cost of validity.* (p.715-716). Nevertheless, most of the focus of the EBGs programme has been on developing better *"strategies to improve compliance with evidence-based clinical management guidelines"* (Frankel et al., 1999. p. 533).

As I shall discuss in subsequent sections, there are at least two ways to measure the quality of healthcare (Donabedian, 1981; 2005): in terms of measures of process or in terms of patient outcomes. As we will see, several advocates of the EBGs movement have focused on what is known as *measures of process* rather than *patient outcomes*. Of course, measuring quality of care in terms of compliance with certain processes has several practical advantages (Goddart et al., 2002). But the main worry with this strategy is that it does not really tell us whether the interventions promoted by EBGs truly improve the health of patients. Rather it assumes that,

[92] Readers interested in what has come to be known as the "quality of care" movement can consult Marjoua and Bozic, 2012 and Luce et al., 1994.

because such interventions are based on the best research available, their successful implementation will automatically bring about improvements in patient outcomes (e.g. Mittal et al., 2014; Bowyer et al., 2014; Pereira et al., 2014).[93]

5.2.3. Basic concepts for this chapter's analysis

In this chapter's analysis, the EBM approach will be equated to an approach where physicians' therapeutic recommendations are in accordance with EBGs. As explained earlier, this approach aims at standardizing care by encouraging compliance with recommendations based on the best research available, and therefore effectively reduces the latitude for the exercise of clinical discretion.[94]

Furthermore, it is important to note from the outset that in this chapter my attention will be restricted to *therapeutic EBGs*. This is not because there are no EBGs developed for diagnostic and prognostic purposes (in fact, there are many of them). The main reason for this is that the (a) therapeutic recommendations and (b) diagnostic and prognostic ones need not be based on the same kind of data (§ 1.3.2 and 1.3.3). Because of this, I shall examine the empirical merits of the EBM approach against the DA separately for therapeutic prescriptions and diagnostic and prognostic predictions.[95] I shall consider therapeutic prescriptions in the present chapter, and turn to diagnostic and prognostic predictions in the next chapter.

Finding a suitable empirical proxy for the DA is, of couse, not a straightforward task. The ideal situation would have been one in which the DA could have been operationalised by measuring the clinical outcomes that followed from therapeutic recommendations made by a random sample of physicians whose judgment was (i) informed by everything they know

[93] The other side of the coin, which we already discussed in chapter 2 (§ 2.4.2), is the assumption that lack of compliance with interventions supported by the best evidence will result in worse patients outcomes (Grimshaw et al., 2004).

[94] Focusing on the performance of EBGs is not necessarily the only satisfactory way to examine the empirical consequences of the EBM approach. In this regard, the application of other models of EBM practice, such as the 4S model (§ 2.4.1), may well result in different patient outcomes. Nonetheless, since both the EBGs model and the 4S model share a commitment to the rules of EBM and consequent marginalization of the PEI, it seems reasonable to assume that these different EBM models of practice will have relatively similar effects on patient outcomes. Still, it is worth recognizing that there may be at least two important sources of variance between the EBGs model and the 4S model. First, there is the time lag with which the best valid research is incorporated into EBGs, which might be shorter in the case of the 4S model, since physicians themselves might be more flexible about keeping themselves updated. Second, since the 4S model, unlike the EBGs model, does not involve official standards of practice, it might provide EBM practitioners with more latitude in adapting their recommendations to their patients' personal utilities, which might be reflected in more individualised recommendations and ultimately in better outcomes. Nonetheless, I do not think that these differences outweigh the commonalities between the EBGs model and the 4S model, and so I shall continue to assume that at least some extrapolations, albeit qualified, are possible.

[95] There are further practical reasons for separating therapeutic from diagnostic and prognostic judgment. The literature comparing the accuracy and of clinical judgment in the context of diagnostic and prognostic tasks (§ 5.2) is highly technical, and requires detailed attention to subtleties. By contrast, comparisons between EBGs and physicians' recommendations in the context of therapy are more straightforward, involving less complex methodological underpinnings and requiring less background knowledge.

about the patient (which includes information from any source the physician deems relevant to the patient, including valid research evidence (§ 1.5.2), and (ii) took into account the patient's preferences and values when aiming to maximize expected utility (§ 1.5.1).

This issue is important because it would inappropriate to devalue the DA by measuring the performance of physicians who are ignorant of crucial research or careless about patients' welfare. The point of the DA is to overcome excusive reliance on controlled research and to allow physicians to address the PEI. The DA is not designed to encourage physicians to go back to the dark pre-EBM ages when valid research played very little role in informing decision-making. So we would like our comparisons to be based on DA practitioners who do not ignore EBGs, but rather consider their content in addition to the extra information they have about the patient.

Regrettably, as this is an empirical investigation based on research already conducted, and as far as the author's knowledge there are no physicians trained in the exercise of the DA, there was no good alternative to assuming that "usual care"–that is, the standard control group in investigations of the impact of EBGs in healthcare, would be a "workable" proxy with sufficient similarity to the DA to permit an informative comparison.

The rationale for using the performance of physicians exercising "usual care" as a proxy for the DA is based on the following assumptions. First, it is reasonable to expect that the impact of clinical practice according to "usual care" will rarely be better than that of physicians' exercising their clinical judgment according to the DA. After all, the performance of physicians practicing "usual care" will be normally hampered by all (or most) the limitations normally attributed to the exercise of clinical judgment (e.g. decisional inconsistency, vulnerability to errors in human reasoning, etc.), as well as a potential neglect of relevant research evidence due to factors such as laziness or carelessness. In fact, it is possible that "usual care" recommendations are made in a generally haphazard way, without seriously addressing the PEI at all.

Given this, if the data indicated that EBGs perform no better that "usual care", this would be highly suggestive that the DA is superior to EBGs. Of course, on the other hand, if the data suggested that EBGs lead to better patient outcomes than "usual care", this would not be a knock out argument against the prospects of the DA, even though it would provide some support for the hypothesis that compliance with EBGs in real settings can improve patients' outcomes.

In the light of these points, this chapter will settle for comparing EBGs with "care as usual". Since my analysis will indicate that EBGs do not in fact lead to better outcomes than "usual

care", this will provide some support for preferring the DA to the EBM approach.

5.2.4. Outcome measures

As mentioned earlier, the performance of different approaches to clinical practice can be studied by both (a) process and (b) outcome measures (Donobedian, 2005, Marjoua and Bozic, 2012). Process measures are focused on the actions performed by clinicians during their practice (Donobedian, 2005). In studies using process measures the benchmark is some intervention or procedure (e.g. ordering of a test, prescription for a medication) that has been a priori specified as normatively adequate and attributed a certain benefit for the patient (Aday et al., 2004).

Two remarks about process measures are in order. First, in the era of EBM, measures of process are standardly defined by paying attention to the best available research evidence, cost-benefit considerations, and applicability analyses.[96] Second, a noteworthy feature about the process of evaluation of EBGs is that many interventions have been considered "successfully implemented" once there is evidence of improvement in process outcomes; that is, when there are data suggesting that an increased number of physicians is complying with the process recommended, regardless of the consequences of those processes on the health status of patients (Ramsdale and Dale, 2013).

However—and following the most recent recommendations for *comparative effectiveness research* and *quality measures in healthcare* (Berlin and Cepeda, 2012; Berger et al., 2012; Chung and Shauver, 2009)–the present analysis will compare EBGs with "usual care" not in terms of process measures, but in terms of patient outcomes. Attention to patient outcomes is essential because of the need to avoid circularity. Using measures of process to assess the benefits of EBGs in is not informative in the present context, precisely because such quality indicators presuppose that the mere act of complying with EBGs improves the care of patients. By contrast, patient outcomes count as independent indicators because they directly measure the patient's health status (Donabedian, 2003, 2005). Patient outcomes include various markers of disease (e.g. blood pressure, body mass index, glycated haemoglobin), and clinical endpoints whose occurrence is considered relevant (e.g. death, disease progression or complications).

It must be noticed, however, that the availability of some measure of *process* is also important to ensure the adequacy comparison between EBGs and "usual care". This is because measures of process provide specific information about physicians' compliance with EBGs. In the

[96] The interested reader is invited to consult Qaseem and colleagues (2010), who describe the process followed in developing EBG for the American Colleague of Physicians (ACP).

absence of such measure, it would be unclear whether the effect on patient outcomes observed in the EBGs group can be truly attributed to the implementation of EBGs, or to the interference (positive or negative) of physicians' judgment who decided, for some reason, not to comply with EBGs.[97]

So in this chapter I shall be examining the relative performance of EBGs (as proxy for the EBM approach) and "care as usual" (as proxy of the DA approach) with respect to patients' outcomes in the context of therapy. And my focus will be on the performance of successfully implemented EBGs, so as to ensure, as much as possible, that we are comparing research-based guidelines with clinical judgment.

5.3. Methods

5.3.1. Search strategy

A systematic literature search was conducted in three electronic databases (*Cochrane Database of Systematic Reviews*, Medline and Embase) to identify *systematic reviews* that studied the impact of EBGs in medical practice. Given that the differences between EBGs and other types of guidelines is not completely clear (Grilli et al., 2000; Shekelle et al., 1999; Wolf et al., 2011, Brouwers et al., 2010), we judged it appropriate to initiate our analysis by adopting a comprehensive search strategy, in which we attempted to find every systematic review assessing the impact of clinical guidelines in medicine.[98]

Databases were searched without language restriction, covering a period from 1950 to March 2015, and using several combinations of keywords (Appendix 1). Systematic reviews of related topics (e.g. health policy, comparative-effectiveness research), references from systematic reviews already retrieved, and other relevant articles (including editorials, commentaries, narrative reviews, and technical reports on EBGs) were examined to identify additional *systematic reviews* missed during the primary search.

Two reviewers independently screened the titles and abstracts, and identified potentially relevant *systematic reviews*. These articles were further selected according to the following

[97] This points to another way in which the worth of EBGs could be overvalued: insofar as we are demanding a measure of process to ensure that EBGs were actually followed, this might selec cases where the reason for compliance with EBGs was precisely because physicians exercised their judgment and concluded that the best recommendation for the patient was the one advised by the EBG. Nonetheless, since the conditions under which physicians have to comply with EBGs in studies, as opposed to real clinical circumstances (Grimshaw, 2004), are relatively fixed, I do not think that the magnitude of this influence is need be a serious concern.

[98] Note, however, that this was only for the sake of completeness: our comparisons required primary studies measuring the impact of EBG recommendations based on assessment of research evidence and not relying exclusively on consensus of experts. On this matter we adopted a similar criteria to Wolf and colleagues (2011), Bahtsevani and colleagues (2004), Barbui and Cipriani (2011), and the collaboration AGREE (Brouwers et al., 2010; Cluzeau et al., 2009). For more details on EBGs, return to section 4.3.1.3 in this chapter.

criteria. Since our main interest was to contrast the EBGs with "usual care" as applied by physicians we do not include *systematic reviews* assessing the impact of EBGs directed to change the practice of other health professionals (nurses, dentists, social workers, etc.), or attempting to promote structural changes in the organization of care that were not directly related to the exercise of physicians' judgment in the context of therapeutic recommendations. In addition, *systematic reviews* investigating the impact of interventions not integrated into EBGs or quality assurance procedures not associated with EBGs were excluded due to the absence of a clear basis on which to establish an adequate comparison between the EBGs and "usual care". Discrepancies between reviewers were resolved by discussion and consensus.

5.3.2. Selection of primary studies

For the reasons presented in (§ 5.3.1.1 and § 5.3.1.2) we only included primary studies in which both process and patient outcomes were considered. In order to minimize the loss of potentially relevant information, and in accordance with the criteria followed by the majority of previous reviewers in this area (Grimshaw, 1993; Worrall et al., 1997; Bahtsevani et al., 2004; Weinman et al., 2007; Lugtenberg et al., 2009), we did not restrict the selection of primary studies to randomized trials [99]. After the exclusion of duplicates each primary study was examined. During this process we excluded all studies (i) in which there were clear problems with the comparability of the groups (obvious differences between the groups with respect to known confounding causes) or (ii) where it was unclear whether physicians in the EBGs group were explicitly authorised to rely on their judgment before making a recommendation in accordance with EBGs. In addition, since our interest was to assess the effect of EBGs on patients' outcomes, we excluded studies in which we were not able to determine that the recommendations were explicitly generated with attention to valid research results (§ 5.3.3 and footnote 6). The full list of primary studies excluded from the qualitative synthesis and the reasons for their exclusion can be found in (Appendix 1).

5.3.3. Data extraction and management of missing data

The following data were collected from each primary study: (1) clinical area (medical condition for which the guideline was designed) (2) study design (RCT, Before and after study, and interrupted time series analysis) (3) participants and setting of the study, (4) type of intervention applied to the intervention and control group, (5) process measures, (6) patient outcomes, (7) effects on process measures, and (8) effects on patient outcomes. Whenever a parameter was not given in a primary study, it was coded as "not provided".

[99] Notice that most types of study designs in this subject, including randomized trials, are subject to the *Hawthorne effect* (the positive effect on the performance of physicians just because they know that their clinical recommendations are being examined) See Grimshaw and Russell (1993).

5.3.4. Quality assessment of primary studies

Two reviewers independently assessed the methodological quality of the primary studies included using the quality criteria developed by the *Cochrane Effective Practice and Organisation of Care Group* (EPOC) (EPOC, 2002). The EPOC quality criteria checklist includes criteria for randomised trials, controlled before and after studies and interrupted time series. Given important methodological differences across studies we were not able to weight their quality using a single measure (EPOC, 2002).

5.3.5. Data synthesis

The impact of EBGs and "care as usual" on patient outcomes was assessed for each primary study by comparing the difference in the outcomes observed in (i) the group where physicians followed EBGs and (ii) the group where physicians followed "care as usual". The set of primary studies included were summarised in the qualitative synthesis (§ 4.4.2).

Due to the high levels of heterogeneity of the primary studies, we did not pool the results in a

meta-analysis (Deeks et al., 2008). Following the approach adopted by Lugtenberg and colleagues (2009) the effect of EBGs and "care as usual" on patients' outcomes for each primary study was summarized in four categories: ***Category (i) "Effective":*** significant effect on all outcomes measured; ***Category (ii) "Mostly effective"***: significant effect on the majority of outcomes measured; ***Category (iii) "Mostly ineffective":*** study failed to demonstrate a significant effect on the majority of outcomes measured; ***Category (iv) "Ineffective"***: no significant effect on any of the outcomes measured.

5.4. Findings

The process of identification and selection of systematic reviews (SRs) and primary studies are shown in **Figure 1**. Eight SRs were identified. [100] These SRs reported the results of 219 primary studies. After excluding duplicates, 149 were reviewed in full. The majority of the primary studies reported in previous SRs (n= 107, 71.8%), could not be included in this analysis because researchers did not measure patient outcomes or because physicians' compliance in recommendations in the EBGs group could not be established. The remaining 32 primary studies (28.2 %) measured process and patient outcomes and were included in the qualitative synthesis. Previous SRs assessing patient outcomes are shown in **Table 1**. Excluded studies and the reasons for exclusion are listed in (Appendix 2).

[100] Lugtenberg et al., 2009; Worrall et al., 1997; Leeds group, 1994; Bahtsevani et al., 2004; Bazian group, 2005; Grimshaw et al., 1993; Barbui et al., 2014; and Weinman et al., 2007.

Table 1: Systematic reviews included in this analysis:

Author (year)	PS reported	PS reported assessing PO (*)	PS assessing PO not reported in previous SRs (**)	PS included
Grimshaw (1993)	59	9 (15.2%)	9 (15.2%)	1
Leeds (1994)	91	14 (15.4%)	4 (14.3%)	0
Worrall (1997)	13	13 (100%)	3 (100%)	0
Bahtsevani (2004)	8	5 (62.5%)	5 (62.5%)	2
Bazian (2005)	5	3 (60%)	3 (60%)	1
Weinman (2007)	18	13 (72.2%)	13 (72.2%)	0
Lugtenberg (2009)	20	8 (40%)	8 (40%)	6
Barbui (2014)	5	2 (40%)	2 (40%)	0
Total	219 †	67†	42 (28.2%)	10

SR= Systematic Reviews, PS= Primary Studies, PO= Patient Outcomes. † Total includes

* % Relative to the total number of PS included in each SR** % relative to new PO included in each SR

5.4.1. Characteristics of primary studies

The characteristics of the ten primary studies included are described in **Table 2** (see below)

5.4.2. Qualitative synthesis

Of the ten primary studies included, four failed to show significant effects on any patient outcomes measured (category iv)[101], four did not show significant effects on the majority of outcomes measured (category iii)[102] and only two[103] showed significant effects on the majority of outcomes (category ii). None of the studies included showed significant effects on all outcomes measured (category i). Nine of the ten studies included showed significant effects on process of care.[104]

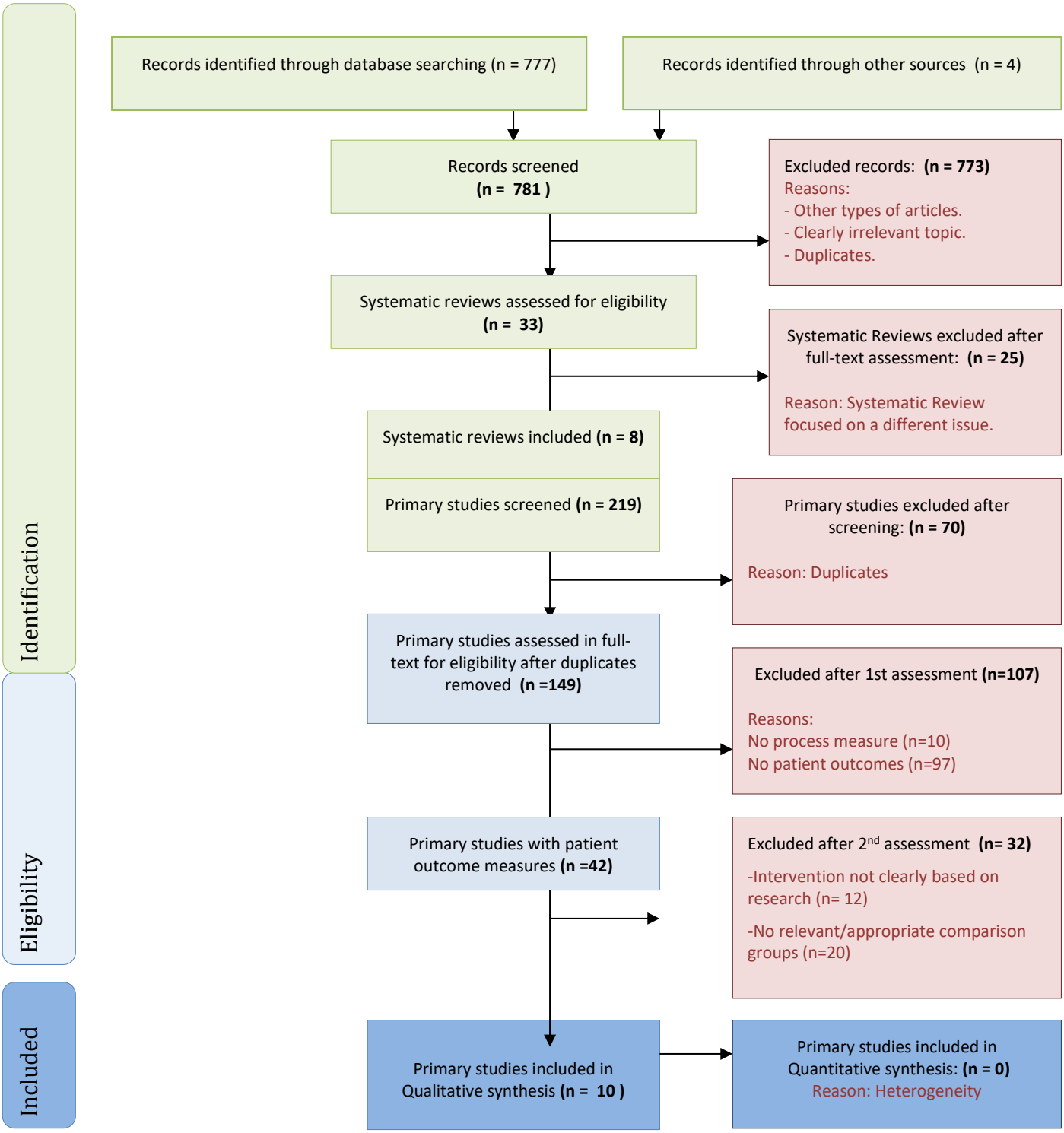
[101] Van Kasteren et al., 2005, Smeele et al.,1999; Renders et al.,2001,2002; and Cummings et al., 1989.

[102] Lobo et al., 2002-2004; Frijling et al., 2002-2003, Jans et al., 2000-2001, Bekkering et al., 2005ab; and Ofman et al., 2003.

[103] Perlstein et al.,1999,2000; and Dufault and Willey- Lessne,1999.

[104] Lobo et al., 2002-2004; Frijling et al., 2002-2003,; Bekkering et al., 2005ab; Ofman et al., 2003; and Cummings et al., 1989; Renders et al., 2001,2002; Jans et al., 2000,2001;Perlstein et al.,1999,2000; and Dufault and Willey-Lessne.,1999; and Van Kasteren et al.,2005.

Figure 1: Flow diagram of study selection process Adapted from PRISMA (*Transparent report of systematic reviews and meta-analyses, 2009.*



Records= Abstracts retrieved. SR= Systematic reviews. PS= Primary studies

Physicians following EBGs failed to show significant improvements in patient outcomes inmost medical conditions studied, including: *Surgical site infections* [105], *Asthma and Chronic Obstructive Pulmonary Disease* [106], *Diabetes type II* [107], *Cardiovascular disease* [108], *Low back pain* [109], *Peptic ulcer* [110], and *Smoking*[111]. In this analysis, physicians who practiced according to "usual care" were outperformed by physicians following EBGs with respect to the inpatient management of *various types of pain* [112] and *Bronchiolitis.* [113]

The overall designs of the studies differ. Five Randomized Controlled Trials (RCTs) [114] and five (non-randomized) before and after studies[115], including one interrupted time series design [116] were included in the qualitative synthesis.

With respect to study designs, four of the five RCTs included indicated improvement in process of care but failed to demonstrate significant improvement in patient outcomes. [117] By contrast, the two studies suggesting that guidelines were more effective than usual care were before and after studies.[118]

5.4.3. Quality analysis

Several methodological aspects of the studies included deserve comment. For a start, in the study conducted by Cummings and colleagues (1989), physicians in the EBGs' group followed recommendations that were based on research evidence published before the formal advent of EBM. This is not necessarily a reason to invalidate a comparison between EBGs and "care as usual", but it must be explicitly acknowledged that what is being examined is the effect on patients' outcomes after the application of recommendations that are currently out-dated.

Furthermore, in the study conducted by Bekkering and colleagues (2005ab) physicians in the

[105] Van Kasteren et al.2005
[106] Smeele et al.1999
[107] Renders et al.2001-2002
[108] Lobo et al. 2002-2004 and Frijling et al. 2002-2003
[109] Bekkering et al.2005ab
[110] Ofman et al. 2003
[111] Cummings et al. 1989
[112] Dufault and Willey- Lessne.1999
[113] Perlstein et al.1999-2000
[114] Lobo et al. 2002-2004; Frijling et al. 2002-2003, Smeele et al.1999; Bekkering et al.2005ab; Ofman et al. 2003; and Cummings et al. 1989.
[115] Renders et al.2001-2002; Jans et al. 2000-2001;Perlstein et al.1999-2000; and Dufault and Willey-Lessne.1999
[116] Van Kasteren et al.2005
[117] Lobo et al. 2002-2004; Frijling et al. 2002-2003; Bekkering et al.2005a and b; Ofman et al. 2003; and Cummings et al. 1989
[118] Perlstein et al.1999-2000; and Dufault and Willey- Lessne.1999.

"care as usual" group were explicitly aware of the recommendations included in EBGs. Although this differs from the standard definition of "usual care", this study was included as closer proxy to the use of clinical judgment in the context of the DA (§ 5.3.1).

In addition, since the study conducted by Smeele and colleagues (1999) physicians' compliance with EBGs was not complete, the lack of effect in patients' outcomes should be interpreted with caution, as part of it could be attributed to non-adherence. Moreover, the results reported by Dufault and Willey- Lessne (1999) with respect to patients' outcomes should also be taken with a grain of salt due to short follow-up periods, large number of dropouts, unclear time frames and a degree of uncertainty as to physicians' compliance in the EBGs group.

As a further matter, although Perlstein and colleagues (1999-2000) reported an improvement in patients' outcomes in the EBGs group, what these authors considered patients' outcomes were indirect measures (e.g. decreased admission rates, length of stay, less resource utilization, etc.), it is unclear whether these outcomes deserved to be classified as such because they could also be a by-product of measures of process which are disconnected with the patients' health status (e.g. a decrease in length of stay could be equally explained by a quicker improvement in the patient's condition or by changes in the process of care such as the introduction of stricter discharge criteria).

Another aspect that deserves attention is that even though the five before and after studies included[119] follow-up periods which were reasonably long, and, more generally, were well-conducted according to the EPOC quality checklist (EPOC 2002), one has to bear in mind that comparisons of populations that are not concurrent in time are particularly susceptible to unknown confounding factors.

Finally, it must be noticed that for the sake of simplicity, we only reported the overall effect size of recommendations on patients' outcomes; however, different recommendations in the EBGs group had different effect sizes (see table 2).

5.5. Discussion

This reanalysis of systematic reviews shows that implementing recommendations supported by valid research evidence via EBGs does not generally improve clinical care measured in terms of patient outcomes. Instead, our analysis indicates that physicians delivering "care as usual" typically achieve better patient outcomes than physicians that comply strictly with EBGs.

[119] Renders et al.2001-2002; Jans et al. 2000-2001;Perlstein et al.1999-2000; and Dufault and Willey-Lessne.1999, Van Kasteren et al.2005.

To our knowledge, this is the most recent and comprehensive reanalysis of systematic reviews comparing the impact of EBGs with "care as usual" and the first one suggesting that the "successful" implementation of EBGs (that is, where evidence-based recommendations where actually adopted by physicians rather than merely providing EBGs to physicians –or making available to them without checking compliance) does not necessarily result in improvements in clinical care over "care as usual".

It is important to stress that this seemingly paradoxical result from the perspective of supporters of EBM, where recommendations backed up by valid evidence did not result in improved patient outcomes, could only be detected because this reanalysis, unlike previous systematic reviews, paid attention to both process and patient outcomes, which permitted us to examine how EBGs work in an ideal scenario – where recommendations were actually followed by physicians in the intervention group.

Notice that our results suggest that sometimes EBGs improve care by a moderate amount compared to usual care. In this respect, it is interesting to speculate as to what would happen with EBGs outside research settings, that is, if EBGs were poorly implemented. On the one hand, one might think that this will improve care even less because deficient implementation would result in physician neglecting potentially relevant recommendations that might help some patients. This line of thinking leads us to believe that since suboptimal implementation is closer to reality and the ideal implementation suggest ineffectiveness, EBGs in the real world are surely not as beneficial to patient outcomes as it is widely supposed. But on the other hand, one might think that suboptimal implementation might bring about unintended positive consequences for patients, for that would imply less pressure for physicians to comply with EBGs and therefore more room to apply personal guidelines based on their judgment informed by everything they know about the patient. However, it must be remembered that, according to the DA, the case for increasing the room for clinical discretion is not substantiated by the idea that judgment alone (that is, judgment uninformed by relevant research) necessarily leads to better recommendations, but rather by the idea that clinical discretion is necessary in the presence of the problem of extra information (PEI).

It is also important to compare the findings of this reanalysis with those of previous systematic reviews. In this regard, our results are at odds with those of two landmark reviews on the effectiveness of EBGs in the United Kingdom (Grimshaw, 1993a; Leeds, 1994), which supported the launch of a structured program to develop EBGs, and a research agenda to investigate how to increase physicians' adherence to EBGs. Both Grimshaw and colleagues (1993a) and the review conducted by the University of Leeds (1994) concluded that *"guidelines do improve clinical practice, when introduced in the context of rigorous*

evaluations."(Grimshaw et al., 1993a. p.1317), and that *"properly developed guidelines can change clinical practice and may change inpatient outcome"* (Leeds, 1994. p.3). However, close examination of their findings revealed that most comparisons were focused on process measures and comparisons focused on patients outcomes where few and inconclusive.

Worrall and colleagues (1997), who focused on EBGs applied in primary care, Bahtsevani and colleagues (2004), and more recently Barbui and colleagues (2014) presented more qualified conclusions. These authors suggested that *"there is very little evidence that the use of [EBGs] improve patients outcomes…"* (Worrall et al., 1997. p.1705), that *"There is a tendency toward support for the idea that outcomes improve for patients…if evidence-based clinical practice guidelines are used, although these findings could be specific to the settings and context of the studies reported in this systematic review."* (Bahtsevani et al., 2004. p.427), and that *"…uncertainty remains in terms of clinically meaningful and sustainable effects of treatment guidelines on patient outcomes…"* (Barbui et al., 2014. p.1).

On the other hand, the Bazian group (2005) acknowledged that the findings were scarce and methodologically problematic, but opted to favour optimistic conclusions *"increasing the production and availability of EBGs can improve quality of care"* (p.270).[120]

Weinman and colleagues (2007) –who focused on psychiatric EBGs– opted to remain largely agnostic as to the efficacy of EBGs on patients' outcomes: *"There is insufficient high-quality evidence to draw firm conclusions on the effects of implementation of specific psychiatric guidelines"* (p.420). However, they explicitly recognised that the *"effects on provider performance or patient outcome were moderate and temporary in most cases"* (p.420). Finally, in consistency with our findings, Lugtenberg and colleagues (2009) confirmed that the Duch experience with EBGs was that *"EBGs can be effective in improving the process and structure of care. [But] the effects of guidelines on patient health outcomes were studied far less and data are less convincing."* (p.385).

So, in summary, while most previous reviewers recognised the lack of evidence confirming the superiority of physicians following EBGs rather than "usual care", they were not able to rule out the possibility that this was due to poor implementations of EBGs—because they failed to focus on patients' outcome in cases where EBGs *were* successfully implemented.

This review has limitations that deserve discussion. As previously mentioned, the small number of studies conducted to compare EBGs with usual care with respect to patients' outcomes, and the poor quality of the primary studies already conducted, introduces

[120] It must be noticed that the Bazian group is not constituted by academic researchers but it is an organization that conducts reviews commissioned by third parties (e.g. pharmaceutical companies).

uncertainty as to the robustness of the findings. Furthermore, another limitation of this analysis is that our original set was constituted by the reviews originally included by previous reviewers, so it might be possible that our conclusions have inherited problems in the methods used to select studies used by previous reviewers. Although it is reasonable to think that eight systematic reviews conducted by different groups provides a sufficiently appropriate universe of primary studies, one might worry about specific reviews on the impact of EBGs on particular diseases, which might have been missed by general systematic reviews. However, an informal search for EBGs published on specific diseases conducted in PUBMED did not reveal new primary studies.

Another worry about the implication of this systematic review is that the standards for developing guidelines have been actively evolving during the last 20 years. This might provide supporters of EBM with some grounds to claim that guidelines that are truly "evidence-based" will usually lead to better patient outcomes. However, this an untested hypothesis, and moreover one that is not supported by the current evidence. Given this, it is striking that guideline makers continue to focus their attention on improving adherence to EBGs, rather than on measuring whether EBGs truly improve patient outcomes.

Overall, then, there seems no good empirical basis for the view that strict compliance with EBGs is general better than "usual care" in the clinical context. Our analysis suggests that the application of such recommendations, in many cases, does not improve patient outcomes. Moreover, since it seems reasonable to take that the performance of "care as usual" as a lower limit to worth of the DA, this reanalysis of published systematic reviews provides some support for the thesis that allowing physicians to apply "personal guidelines" would improve patient outcomes.

Table 2: Primary studies measuring the performance of guideline recommendations against usual care.

Study	Clinical area	Study design	Participants / Setting	Guideline group (GG)	Control group (CG)	Effect on process of care	Effect on patient outcomes
Van Kasteren et al. (2005)	Surgical site infections	Before & after.(Interrupted time series)	13 hospitals; 1763 procedures before/ 2050 after.	Implementation of guideline that included performance feedback.	Before guideline	**Effective:** Improvement in all four measures of process.	**Ineffective:** No changes on overall Surgical Site Infections rates
Smeele et al. (1999).	Asthma & COPD	RCT (Cluster).	34 GPs were randomized: GG (n=17) & CG (n=17). 433 patients: GG (n= 210) & CG (n=223).	Implementation of guideline: group education and peer review programme.	No intervention	**Mostly ineffective:** Improvement in two measures of structure of care. No change on any of the six measures of process of care. **Notice lack of adherence to process.	**Ineffective:** No improvement in any of the 3 patient outcomes (symptoms, smoking habit, disease specific quality of life) measured.
Renders et al. (2001&2002)	Diabetes type II	Before and after study (controlled).	27 GPs: GG (n= 22) & CG (n= 5). 389 patients: GG (n= 312) & CG: (n= 77).	Implementation of guideline: distribution, education, audit and feedback.	Before guideline	**Effective:** Improvement in all 9 indicators of process.	**Ineffective:** No improvement in any of the 14 patient outcomes measured (BP, HbA1c, etc.).
Lobo et al. (2002 & 2004) & Frijling et al. (2002 & 2003)	Cardiovascular disease	RCT (Cluster)	124 Practices were randomized. 185 GPs. 2268 patients: Diabetes (n= 537), Cardiovascular Disease (n= 617), and Hypertension (n= 1114).	Implementation of guideline: support from facilitators, discussion of feedback reports, and evaluation during outreach visits.	No intervention	**Mostly effective:** Improvement in process of cardiovascular care (5 of 12 indicators) and process of diabetes care (2 of 7 indicators).	**Mostly ineffective:** No improvement in any aspect in patients with hypertension. Improvement in only 2 of 8 aspects of HRQL in diabetes patients and in only 3 of 8 aspects in patients with cardiovascular disease.
Jans et al. (2000 & 2001)	Asthma & COPD	Before and after study (Controlled)	19 practices: GG (n=14) & CG (n= 5). 370 patients: GG (n= 280) & CG (n=90).	Implementation of guideline: identification of barriers, documentation of the care provided, specific education, feedback and peer review.	Before guideline	**Mostly effective:** Improvement in 4 of 8 aspects (including monitoring of medication compliance and measurement of PEFR).	**Mostly ineffective:** Improvement in only 1 of 4 outcomes measured (PEFR).

Bekkering et al. (2005a & b)	Low back pain	RCT (Cluster)	Practices were randomised. Block-randomisation: Physiotherapists (n= 113); practices (n= 68).	Implementation of guideline: Dissemination and active training strategy (education, discussion, role playing, feedback and reminders).	Standard dissemination	**Mostly effective:** Improvement in all 4 measures of process (including: setting functional treatment goals and giving patient information).	**Mostly ineffective:** No significant difference in outcomes (physical functioning and pain) despite the difference in process measures between groups.
Perlstein et al. (1999 & 2000)	Bronchiolitis	Before and after study (Controlled)	Records (n=1.979) from guideline-eligible patients were reviewed.	Guideline implementation for the inpatient care of infants with bronchiolitis.	No intervention	**Mostly effective:** Improvement in 5 measures of process (including, guideline use, use of antibiotics, use of blood gas and use of B2 agonists)	**Mostly effective:** Decrease in admission rates, readmission rates and lengths of stay in hospital
Dufault & Willey-Lessne. (1999)	Pain management	Before and after study	Clinicians (n=102) Surgical and oncology patients with a history of pain in the month before hospitalization (n = 239) *** CG not well described.	Use of clinical pathways in management of cancer pain, acute and traumatic pain, and low back pain.	Before guideline	**Mostly effective:** Clinicians adherence to clinical pathways was checked by clinical audit.	**Mostly effective:** Improvement in 4 patient outcomes measured at three days but no difference when patients were discharged home.
Ofman et al. (2003)	Peptic ulcer	RCT (Cluster)	8 primary care clinics, 406 patients. GG (n= 200) CG (n=206)	Guideline implementation for the management of dyspepsia and eradication of Helicobacter pylori (Education and provision of serological testing)	No intervention	**Effective:** Improvement in all 3 process measures (H pylory testing, use of recommended regimen, Proton-pump inhibitor discontinued after 8-12 weeks)	**Mostly ineffective:** No improvement health-related quality of life and satisfaction with care at 6 months. No significant difference in 15 of 16 symptom scores.
Cummings et al. (1989)	Smoking cessation	RCT	4 Medical centres. 81 Physicians: GG (n=40) CG (n=41)	3 hours of training. Self-help booklets to distribute free to smokers, reminders.	No intervention	**Effective:** Improvement in 5 process measures (including more counselling and more appointments)	**Ineffective:** No significant improvement in rate of abstinence after 9 months

Effects of Guideline Recommendations on patient outcomes (10 Studies from 17 Articles.). **Categories: Effective** (significant effect on all outcomes), **Mostly effective** (significant effect on the majority of outcomes), **Mostly ineffective** (no significant effect in the majority of outcomes), **Ineffective** (no significant effect on any outcome).

5.6. Conclusions

Overall, the results of this reanalysis of systematic reviews comparing EBGs and "care as usual" with respect to patient outcomes provides no evidence that physicians' adherence to EBGs leads to superior clinical care. In the majority of the studies, compliance with EBGs was reflected in improvements in measures of process, but this did not translate into better outcomes for patients. Insofar the comparison between the EBGs and "usual care" constitutes a proxy for the relative performance of the EBM approach the DA, this reanalysis challenges the view that the EBM approach is superior to the DA. Furthermore, our findings address some concerns about the possible negative consequences of allowing physicians to exercise their clinical discretion, and thereby vindicate, albeit partially, the viability of the DA in the context of therapy. Nonetheless, due to the limitations presents in primary studies, further studies specifically designed to test the performance of EBGs against the DA are needed. Finally, since the EBGs model is only one way to exercise EBM, a full vindication of the DA in the context of therapy would require additional studies examining the relative performance of other versions of EBM (such as the 4S model) against the DA.

5.7. Appendices

5.7.1. Search strategy

Search strategy to identify Systematic Reviews on the effectiveness of guidelines

Keywords:

1. exp Practice Guidelines as Topic/
2. guideline.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
17. 1 or 2
3. exp Treatment Outcome/
14. exp "Outcome and Process Assessment (Health Care)"/
4. outcome.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
7. patient.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
18. 4 and 7
10. effectiveness.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
11. clinical.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
19. 10 and 11
21. 3 or 14 or 18 or 19
9. exp Comparative Effectiveness Research/
50. 9 or 21
15. exp "Process Assessment (Health Care)"/
16. process.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier]
20. 4 and 16
22. 15 or 20
51. 17 and 22 and 50
52. limit 51 to systematic reviews

777 total

5.8. References

Aday, L.N., Begley, C.E. and Lairson, D.R. (2004). *Evaluating the Healthcare System: Effectiveness, Efficiency, and Equity. 3rd Ed.* Chicago, The Foundation of the American College of Healthcare Executives.

Bahtsevani, C., Udén, G. and Willman, A. (2004). Outcomes of evidence-based clinical practice guidelines: a systematic review. *Int J Technol Assess Health Care.* 20(4):427-33.

Barbui, C. and Cipriani, A. (2011). What are evidence-based treatment recommendations? *Epidemiol Psychiatr Sci.* 20(1):29-31.

Barbui, C., Girlanda, F., Ay, E., Cipriani, A., Becker, T. and Koesters, M. (2014). Implementation of treatment guidelines for specialist mental health care. *Cochrane Database Syst Rev.* (1):CD009780.

Bazian Ltd. (2005). Do evidence-based guidelines improve the quality of care? *Evidence-Based Healthcare & Public Health*. 9:270-5.

Bekkering, G.E., Hendriks, H.J., van Tulder, M.W., Knol, D.L., Hoeijenbos, M., Oostendorp, R.A. and Bouter, L.M. (2005). Effect on the process of care of an active strategy to implement clinical guidelines on physiotherapy for low back pain: a cluster randomised controlled trial. *Qual Saf Health Care.* 14(2):107-12.

Bekkering, G.E., van Tulder, M.W., Hendriks, E.J., Koopmanschap, M.A., Knol, D.L., Bouter, L.M. and Oostendorp, R.A. (2005). Implementation of clinical guidelines on physical therapy for patients with low back pain: randomized trial comparing patient outcomes after a standard and active implementation strategy. *Phys Ther.* 85(6):544-55.

Berger, M.L., Dreyer, N., Anderson, F., Towse, A., Sedrakyan, A. and Normand, S.L. (2012). Prospective observational studies to assess comparative effectiveness: the ISPOR good research practices task force report. *Value Health.* 15(2):217-30.

Berlin, J.A. and Cepeda, M.S. (2012). Some methodological points to consider when performing systematic reviews in comparative effectiveness research. *Clin Trials*. 9(1):27-34.

Bosse, G., Breuer, J.P. and Spies, C. (2006). The resistance to changing guidelines--what are the challenges and how to meet them. *Best Pract Res Clin Anaesthesiol*. 20(3):379-95.

Bowyer, A., Jakobsson, J., Ljungqvist, O. and Royse, C. (2014). A review of the scope and measurement of postoperative quality of recovery. *Anaesthesia*. 69(11):1266-78.

Brouwers, M.C., Kho, M.E., Browman, G.P., Burgers, J.S., Cluzeau, F., Feder, G., Fervers, B., Graham, I.D., Grimshaw, J., Hanna, S.E., Littlejohns, P., Makarski, J. and Zitzelsberger L; AGREE Next Steps Consortium. (2010). AGREE II: advancing guideline development, reporting and evaluation in health care. *CMAJ.* 182(18):E839-42.

Chung, K.C. and Shauver, M.J. (2009). Measuring quality in health care and its implications for pay-for-performance initiatives. *Hand Clin.* 25(1):71-81, vii.

Cluzeau, F. (2009). Conflicting recommendations. Let's not forget AGREE. *BMJ.* 338:b407.

Cummings, S.R., Coates, T.J., Richard, R.J., Hansen, B., Zahnd, E.G., VanderMartin, R., Duncan, C., Gerbert, B., Martin, A. and Stein, M.J. (1989). Training physicians in counseling about smoking cessation. A randomized trial of the "Quit for Life" program. *Ann Intern Med*. 110(8):640-7.

Deeks, J.J., Higgins, J.P. and Altman, D.G. (2008). *Analysing Data and Undertaking Meta-Analyses.* In *Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book* Series. J.P. Higgins and S. Green (eds). Chichester, John Wiley & Sons Ltd.

Donabedian, A. (1981). Criteria, norms and standards of quality: what do they mean? *Am J Public Health.* 71(4):409-12.

Donabedian, A. (2003). *An introduction to quality assurance in health care*. New York, Oxford University Press.

Donabedian, A. (2005). Evaluating the Quality of Medical Care. *Milbank Q.* 83(4):691-729.

Dufault, M.A. and Willey-Lessne, C. (1999). Using a collaborative research utilization model to develop and test the effects of clinical pathways for pain management. *J Nurs Care Qual.* 13(4):19-33.

EBM Working Group. (1992). Evidence-Based Medicine: a new approach to teaching the practice of medicine. *JAMA.* 268(17):2420-5.

Eddy, D.M. (1982). *Probabilistic reasoning in clinical medicine: problems and opportunities*. In *Judgement under uncertainty: Heuristics and Biases.* D. Kahneman and A. Tversky (eds). Cambridge, Cambridge University Press. pp. 249–67.

Eddy, D.M. (1990a). The Challenge. *JAMA.* 263(2):287-90.

Eddy, D.M. (1990b). Clinical decision making: from theory to practice. Practice policies --what are they? *JAMA.* 263(6):877-8, 880.

Eddy, D.M. (1990c). Practice Policies: Where Do They Come From? *JAMA.* 263(9):1265, 1269, 1272.

Eddy, D.M. (1990d). Clinical decision making: from theory to practice. Practice policies--guidelines for methods. *JAMA.* 263(13):1839-41.

Eddy, D.M. (1990e). Clinical decision making: from theory to practice. Guidelines for policy statements: the explicit approach. *JAMA.* 263(16):2239-40, 2243.

Eddy, D.M. (1990f). Clinical decision making: from theory to practice. Designing a practice policy. Standards, guidelines, and options. *JAMA.* 263(22):3077, 3081, 3084.

Eddy, D.M. (2005). Evidence-based medicine: a unified approach. *Health Aff (Millwood).* 24(1):9-17.

Eddy, D.M. (2011). The origins of evidence-based medicine--a personal perspective. *Virtual Mentor.* 13(1):55-60.

EPOC. (2002). *Data Collection Checklist.* In *EPOC Resources for review authors.* Oslo, Norwegian Knowledge Centre for the Health Services.

Frijling, B.D., Lobo, C.M., Hulscher, M.E., Akkermans, R.P., Braspenning, J.C., Prins, A., van der Wouden, J.C. and Grol, R.P. (2002). Multifaceted support to improve clinical decision making in diabetes care: a randomized controlled trial in general practice. *Diabet Med.* 19(10):836-42.

Frijling, B.D., Lobo, C.M., Hulscher, M.E., Akkermans, R.P., van Drenth, B.B., Prins, A., van der Wouden, J.C. and Grol, R.P. (2003). Intensive support to improve clinical decision making in cardiovascular care: a randomised controlled trial in general practice. *Qual Saf Health Care.* 12(3):181-7.

Grilli, R., Magrini, N., Penna, A., Mura, G. and Liberati, A. (2010). Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet.* 355(9198):103-6.

Grimshaw, J.M. and Russell, I.T. (1993). Achieving health gain through clinical guidelines. I: Developing scientifically valid guidelines. *Qual Health Care.* 2(4):243-8.

Grimshaw, J.M., Thomas, R.E., MacLennan, G., Fraser, C., Ramsay, C.R., Vale, L., Whitty, P., Eccles, M.P., Matowe, L., Shirran, L., Wensing, M., Dijkstra, R. and Donaldson, C. (2004). Effectiveness and efficiency of guideline dissemination and implementation strategies. *Health Technol Assess.* 8(6):iii-iv, 1-72.

Jans, M.P., Schellevis, F.G., Van Hensbergen, W. and van Eijk, J.T. (2000). Improving general practice care of patients with asthma or chronic obstructive pulmonary disease: evaluation of a quality system. *Eff Clin Pract.* 3(1):16-24.

Jans, M.P., Schellevis, F.G., Le Coq, E.M., Bezemer, P.D. and van Eijk, J.T. (2001). Health outcomes of asthma and COPD patients: the evaluation of a project to implement guidelines in general practice. *Int J Qual Health Care.* 13(1):17-25.

Klein, J.G. (2005). Five pitfalls in decisions about diagnosis and prescribing. *BMJ.* 330(7494):781-3.

Lacchetti, C. and Guyatt, G. (2002). *Surprising results of randomized controlled trials*. In *Users' guide to the medical literature.* G.H. Guyatt and D. Rennie (eds). Chicago, AMA Press.

Leeds group. (1994). Implementing clinical practice guidelines. Can guidelines be used to improve clinical practice? *Efficient health care bulletin*. 8.

Lobo, C.M., Frijling, B.D., Hulscher, M.E., Bernsen, R.M., Braspenning, J.C., Grol, R.P., Prins, A. and van der Wouden, J.C. (2002). Improving quality of organizing cardiovascular preventive care in general practice by outreach visitors: a randomized controlled trial. *Prev Med.* 35(5):422-9.

Lobo, C.M., Frijling, B.D., Hulscher, M.E., Bernsen, R.M., Grol, R.P., Prins, A. and van der Wouden, J.C. (2004). Effect of a comprehensive intervention program targeting general practice staff on quality of life in patients at high cardiovascular risk: a randomized controlled trial. *Qual Life Res.* 13(1):73-80.

Lugtenberg, M., Burgers, J.S. and Westert, G.P. (2009). Effects of evidence-based clinical practice guidelines on quality of care: a systematic review. *Qual Saf Health Care.* 18(5):385-92.

Marjoua, Y. and Bozic, K.J. (2012). Brief history of quality movement in US healthcare. *Curr Rev Musculoskelet Med.* 5(4):265-73.

Mittal, V., Hall, M., Morse, R., Wilson, K.M., Mussman, G., Hain, P., Montalbano, A., Parikh, K., Mahant, S. and Shah, S.S. (2014). Impact of inpatient bronchiolitis clinical practice guideline implementation on testing and treatment. *J Pediatr.* 165(3):570-6.e3.

Ofman, J.J., Segal, R., Russell, W.L., Cook, D.J., Sandhu, M., Maue, S.K., Lowenstein, E.H., Pourfarzib, R., Blanchette, E., Ellrodt, G. and Weingarten, S.R. (2003). A randomized trial of an acid-peptic disease management program in a managed care environment. *Am J Manag Care.* 9(6):425-33.

Pereira, S., Hassler, S., Hamek, S., Boog, C., Leroy, N., Beuscart-Zéphir, M.C., Favre, M., Venot, A., Duclos, C. and Lamy, J.B. (2014). Improving access to clinical practice guidelines with an interactive graphical interface using an iconic language. *BMC Med Inform Decis Mak.* 14:77.

Perlstein, P.H., Kotagal, U.R., Bolling, C., Steele, R., Schoettker, P.J., Atherton, H.D. and Farrell, M.K. (1999). Evaluation of an evidence-based guideline for bronchiolitis. *Pediatrics.* 104(6):1334-41.

Perlstein, P.H., Kotagal, U.R., Schoettker, P.J., Atherton, H.D., Farrell, M.K., Gerhardt, W.E. and Alfaro, M.P. (2000). Sustaining the implementation of an evidence-based guideline for bronchiolitis. *Arch Pediatr Adolesc Med.* 154(10):1001-7.

Qaseem, A., Snow, V., Owens, D.K. and Shekelle, P.; Clinical Guidelines Committee of the American College of Physicians. (2010). The development of clinical practice guidelines and guidance statements of the American College of Physicians: summary of methods. *Ann Intern Med.* 153(3):194-9.

Ramsdale, E. and Dale, W. (2013). Evidence-based guidelines and quality measures in the care of older adults. *Virtual Mentor.* 15(1):51-5

Renders, C.M., Valk, G.D., Franse, L.V., Schellevis, F.G., van Eijk, J.T. and van der Wal, G. (2001). Long-term effectiveness of a quality improvement program for patients with type 2 diabetes in general practice. *Diabetes Care.* 24(8):1365-70.

Renders, C.M., Valk, G.D., Van de Poll-Franse LV, Schellevis, F., van Eijk, J. and van der Wal, G. (2002). (in Dutch) De effectiviteit van een kwaliteitsbevorderingsprogramma op de zorg voor diabetespati¨enten in de eerste lijn [Effectiveness of a quality improvement program on care of diabetes patients in primary care]. *Huisarts Wet.* 45:512-7.

Shekelle, P.G., Woolf, S.H., Eccles, M. and Grimshaw, J. (1999). Developing clinical guidelines. *West J Med.* 170(6):348-51.

Smeele, I.J., Grol, R.P., van Schayck, C.P., van den Bosch, W.J., van den Hoogen, H.J. and Muris, J.W. (1999). Can small group education and peer review improve care for patients with asthma/chronic obstructive pulmonary disease? *Qual Health Care.* 8(2):92-8.

Swinglehurst, D. (2005) Evidence-based guidelines: the theory and the practice. *Evidence-Based Healthcare and Public Health*. 9(4):308-14.

Thomas, K.B. (1978). The consultation and the therapeutic illusion. *Br Med J*. 1:1327-8.

van Kasteren, M.E., Mannien, J., Kullberg, B.J., de Boer, A.S., Nagelkerke, N.J., Ridderhof, M., Wille, J.C. and Gyssens, I.C. (2005). Quality improvement of surgical prophylaxis in Dutch hospitals: evaluation of a multi-site intervention by time series analysis. *J Antimicrob Chemother.* 56(6):1094-102.

Weinmann, S., Koesters, M. and Becker, T. (2007). Effects of implementation of psychiatric guidelines on provider performance and patient outcome: systematic review. *Acta Psychiatr Scand.* 115(6):420-33.

Wennberg, J.E. (2002). Unwarranted variations in healthcare delivery: implications for academic medical centres. *BMJ*. 325(7370):961-4.

Wennberg, E. (2011). Time to tackle unwarranted variations in practice. *BMJ*. 342:d1513.

Wolf, J.S. Jr, Hubbard, H., Faraday, M.M. and Forrest, J.B. (2011). Clinical practice guidelines to inform evidence-based clinical practice. *World J Urol.* 29(3):303-9.

Worrall, G., Chaulk, P. and Freake, D. (1997). The effects of clinical practice guidelines on patient outcomes in primary care: a systematic review. *CMAJ*. 156(12):1705-12.

Chapter 6: Clinical versus statistical prediction in diagnosis and prognosis

6.1. Abstract

The central aim of this chapter is to address the concern that the application of the discretionary approach (DA) in the context of diagnosis and prognosis might worsen clinical care. I shall do this by conducting a systematic review of studies comparing the relative predictive performance of "*statistical models*" against physicians' judgment.

Three medical databases were searched up to September 5, 2015. All types of studies comparing the performance of predictions made by validated models against predictions made by physicians under real clinical circumstances were considered. Eligible studies included a measure of discriminative accuracy. Two independent authors selected eligible studies, extracted data, and assessed study risk of bias.

Sixty-eight studies comprising 121 pairwise comparisons were eligible. The accuracy of statistical models and physicians varied importantly across predictive tasks. Differences in discriminative accuracy, measured either by areas under receiver-operating characteristic curves (AUCs) or by overall accuracies (OAs), could be ascertained in 69 comparisons, of which 18 (26·1%) favoured models, 32 (46·4%) favoured physicians, and 19 (27·5%) indicated no significant differences. These proportions allow rejection of the hypothesis that most comparative studies show statistical models to be significantly better than physicians' judgement ($p = 0.000118$).

Available evidence indicates that, by and large, physician' judgment achieves greater predictive discrimination than statistical models. This systematic review thus provides support for the thesis that increasing clinical discretion in response to PEI would not worsen the quality of clinical prediction.

6.2. Background

Sound clinical care rests upon accurate predictions. Yet in many areas of medicine correct diagnosis or prognosis is exceedingly difficult (Shojania et al., 2003; Christakis and Iwashyna, 1998; Hunter, 1996). Nowadays, in part as a result of the influence of the EBM approach to clinical care (Guyatt et al., 2008; Howick, 2011), risk scores, prediction rules and other tools based on statistical modelling have a prominent place in the *armamentarium* with which physicians attempt to deal with clinical uncertainty (Richardson et al., 2015).

Decisions based on statistical tools have several potential advantages. For a start, they offer increased objectivity, standardization of practices, and cost containment (Steyerberg, 2009; Singh et al., 2008). But more fundamentally, as discussed in chapter 4 (§ 4.4) the rationale behind EBM's support for statistical methods of prediction is supposed to be that they typically deliver more accurate predictions than physicians' informal judgment (Howick, 2011). A long research tradition in psychology, dating back to Paul Meehl (Meehl, 1954) and continuing with more recent reviews (Dawes et al., 1989; Marchese, 1992; Grove et al., 2000; Ægisdóttir et al., 2006), seems to support the claim that various types of statistical models generally do better than physicians in a wide range of predictive tasks. However, as previously argued (§ 4.4.3), the validity and relevance of this research to clinical medicine remains unclear. Methodological problems, which include comparisons in derivation samples, use of non-medical judges, reliance on hypothetical cases rather than real patients, and the study of outcomes not pertaining to clinical medicine, make it difficult to know whether statistical models genuinely outperform clinical judgement in actual clinical settings or vice versa.

I also noted in chapter 4 (§ 4.4.2) the use of models is often defended by reference to the extensive literature on the general fallibility of human judgment (Gilovich et al., 2003; Dawson and Arkes 1987; Elstein, 1999), together with studies revealing that physicians exhibit poor forecasting performance with respect to several predictive tasks (Brokaw et al., 2004; Smith et al., 2002; Gerestein et al., 2009; Leung et al., 2012). However, the implications of such biases in clinical care remain unclear.

Moreover, it is noteworthy that many supporters of EBM have overlooked that there are familiar dangers in reliance on models (Howick, 2011).[121] First, a significant number of

[121] Note that not all supporters of EBM have the same opinion as to the merits of statistical models. For example, Knottnerus et al., 2002, in his book "The Evidence Based of clinical Diagnosis", offer what in my opinion is a reasonably balanced view about the merits and limitations of both physicians' judgments and statistical prediction methods.

statistical tools do not meet current methodological standards (Collins et al., 2014; Maguire et al. 2011, Keogh et al., 2014; Mallett et al., 2010); second, several replication studies suggest that the stability of the accuracy of prediction rules across time and geographical areas often lacks robustness (Yap et al., 2006; Bleeker et al., 2003; Crowe et al., 2003; Kong et al., 2014; Rabin et al., 2014); and third, there is always the risk that models will ignore predictively relevant factors that are identifiable in specific clinical scenarios (Laupacis et al., 1997; Sniderman et al., 2015; Van Calster et al., 2015).

In this systematic review we aim to examine studies comparing the relative predictive performance of clinical judgement and statistical models in real clinical settings, so to address the objection that the application of the DA in the context of diagnosis and prognosis might worsen clinical care. Since the previous chapter provide relevant empirical evidence on the performance of therapeutic judgment, the present chapter will focus specifically on predictive accuracy in diagnostic and prognostic tasks. In order to refine our understanding of the potential of the DA, we also examined whether the comparative accuracy of models and medical judgement is affected by physicians' level of training and by whether the models are local rather than imported.

6.3. Methods

6.3.1. Search strategy

We systematically searched MEDLINE, EMBASE and PsycINFO for articles in any language from the earliest date of each database to April 20, 2015 using the Ovid Platform. The search was updated in September 5, 2015. Different search strategies were piloted attending to general guidelines (Sampson et al., 2009; Hausner et al., 2015), and specific recommendations for diagnostic (Beynon et al., 2013) and prognostic studies (Dretzke et al., 2014). The final search was designed using relevant keywords and MeSH terms (§ 6.7.1). Complete reference lists of relevant articles were checked for additional studies.

6.3.2. Eligibility criteria

All type of studies comparing the performance of predictions made by validated models against predictions made by medical physicians under real clinical circumstances were eligible.

To avoid biased comparisons the accuracies of both physicians and models had to be obtained in independent samples from those used during the derivation of the prediction method under comparison (Steyerberg, 2013). Clinical Prediction Rules, Risk scores, Diagnostic Decision Support Systems and similar tools used for diagnostic or prognostic prediction were

admissible. Model-based computer aids to assist the interpretation of psychological and medical tests (e.g. Rorschach test, neuropsychological batteries, electrocardiography, images, etc.) were outside the scope of this study.

Physicians in any stage of training (but not medical students) had to be responsible for making the predictions. Physicians had to have direct access to patients and other sources of information normally available in clinical settings depending on the predictive task. Access to structured data-collection forms was admissible only if this was not the only source of information and the estimates provided by statistical models were not made available to physicians by researchers. Studies where the information available was artificially restricted to pre-selected sets of relevant data (e.g. case summaries) were not included.

The predictive tasks were required to forecast clinical outcomes. Dichotomous, ordinal or nominal clinical outcomes were predicted probabilistically (via point probability estimates, probability intervals, or ordinal risk categories) or binarily (present/absent). In tasks where predictions were closely connected to decisions (e.g. prediction of appendicitis and decision to take the patient to the operating room), physicians' judgments had to be requested explicitly and in a manner directly related to the outcome predicted so to avoid conflation between predictive accuracy and utility assessments affecting clinical decisions.

Studies were required to report accuracy in terms of discrimination measures for one or more thresholds. The report of calibration measures was not necessary for inclusion. Research comparing the adequacy of predictions, subsequent clinical decisions, or both solely in terms of economic costs and other utilities were not included.

6.3.3. Data extraction and quality assessment

Two of the team members were responsible for the selection of potentially relevant studies. Two independent authors assessed the selected full text articles for inclusion, and extracted data on setting, predictive task (medical condition, timing of the prediction and type of outcome), reference standard, mode of elicitation of physicians' predictions (probabilistic or binary), number of physicians and specialty, type of statistical model, population, and place in which models were derived and validated, using a predesigned standardized form. We categorized physicians' experience as (i) physicians in training only, (ii) mixed experience (in training and fully trained physicians), and (iii) fully trained physicians only. Disagreements during data extraction were resolved through referral to a third investigator, and team discussion. The quality of included studies was assessed independently by two authors using criteria selected from pertinent guidelines: QUADAS-2 (Whittng et al., 2011), STARD

(Bossuyt et al., 2003)[122], TRIPOD (Moons et al., 2015), and GRADE statements (Guyatt et al., 2011). The form used to assess the quality of studies included design and type of enrolment, adequacy of the reference standard, risk of verification and incorporation bias, missing data, and quality of statistical analysis and reporting.

6.3.4. Statistical analysis

The main unit of analysis were independent pairwise comparisons. To be considered independent, comparisons had to involve a new population, predictive task, sample of physicians, or statistical model.

Due to the high levels of heterogeneity and the variety of measures reported we did not pool the total set of comparisons. Instead we focused on the direction of findings and tested the hypotheses that statistical models discriminate better than physicians and vice-versa by adopting a method based on the "sign test" (Borenstein et al., 2009).

The accuracies of probabilistic predictions were preferably summarized in terms of single measures averaged across multiple thresholds (areas under received operative characteristic curves (AUCs) and Ordinal c-indices). When these measures were not reported and could not be calculated (e.g. when the predictive task consisted of placing the patient into several nominal categories), the proportion of correct predictions over total predictions (overall accuracy (OA)) was used as the primary measure. When possible, sensitivities, specificities and odds ratios were also reported. We used the cut-off provided by the authors, and when unavailable or multiple cut-offs were possible we used the highest *Youden's index* to select one. *Hosmer–Lemeshow goodness-of-fit* test and standard calibration plots were accepted as measures of calibration.

Standard significance tests for non-independent samples (p-value level ≤ 0.05) or the absence of overlap between respective 95% confidence intervals (CIs) were used as criteria to establish superior discrimination. When unavailable, binomial-proportion CIs were calculated. Comparisons without adequate tests and with overlap between CIs were classified as "inconclusive". The proportion of comparisons favouring one approach was then compared with the proportion of comparisons indicating no difference or favouring the other approach using binomial exact tests at probability 0.5 (Greenland et al., 2008). In addition, for the purposes of maximizing data analysed an inductive threshold ($\leq$ 25% overlap between 95% CIs) was applied to reclassify inconclusive comparisons into those suggesting a difference

[122] The full list of articles that detail the standards for the reporting of diagnostic accuracy studies (STARD) are available in the following address: www.stard-statement.org (accessed 8 January 2015)

between models and physicians and those suggesting no difference.

In the subgroup analyses, the influence of different levels of training and of local versus external derivation were investigated using the *Cochran–Armitage test for trend* and *Fisher's exact tes*t, respectively. Due to the wide range of outcomes and populations included, funnel plots were not used to assess publication bias. This study is registered with PROSPERO, number CRD42015020607.

Supplementary analyses in which we pooled comparisons involving similar predictive tasks and excluded overrepresented outcomes are shown in (§ 6.7.2). Analyses were conducted in Stata version 12 software. Depending on the measures and the number of comparisons we used the commands metandi, metaprop or metan.

6.4. Findings

6.4.1. Study characteristics

We examined 5069 titles and abstracts and selected 225 articles for full-text assessment (Fig. 1). Sixty-eight studies comprising 121 comparisons involving 74 samples of physicians and 79 models met our inclusion criteria. Sixty-two comparisons (51.2%) were focused on diagnostic outcomes, which ranged from ankle fracture to various gastrointestinal conditions such as acute appendicitis. Among the 59 prognostic comparisons included, prediction of mortality was the most common task. Additional prognostic outcomes included complications (e.g. bleeding after myocardial infarction) and condition-specific adverse outcomes such as self-harm. Comparisons were carried out in different settings of care (Table 1). Patient sample sizes varied considerably (median 405, interquartile range (IQR): 231-956). Both adults and children were included in patient samples. The timeframe between prognostic predictions and outcomes ranged from few days to 5 years.

Most studies used one sample of physicians. Only 3 studies reported estimates for more than one sample of physicians. Physician sample sizes were reported in only 43 comparisons (35.5%). When reported the median was 15 physicians (IQR: 9-31). Information about the training level could be obtained for 102 comparisons, of which 18 included only physicians in training, 45 fully trained physicians exclusively, and 39 physicians in both categories. Eighty-seven comparisons (71.9%) were between physicians and models that had initially been derived from patients from different centres (imported models).

Areas under receiver operating characteristic curves (AUCs) were obtained for 59 comparisons from 26 studies. In 28 comparisons (10 studies) AUCs were the only measure

available. Overall accuracies (OAs) were available for 86 comparisons from 56 studies. In 14 comparisons OAs were the only measures available because predictive tasks consisted of deterministic predictions over sets of nominal categories (e.g. differential diagnoses).

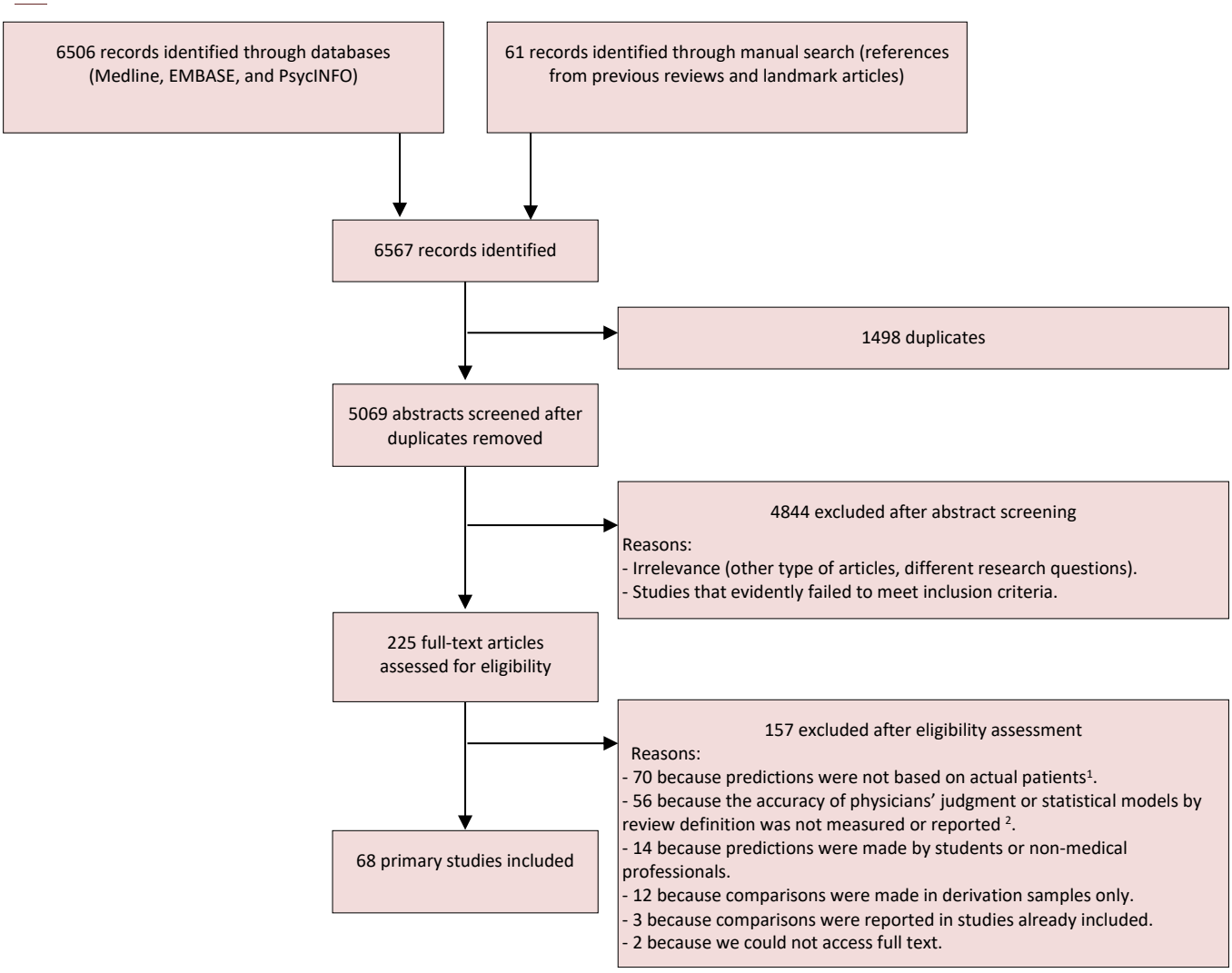
Table 1 presents the main features of the set of comparisons included. The characteristics of each comparison included can be found in (§ 6.7.3).

Table 1: Sample description

Sample description		Comparisons *n* (%)
Comparison settings †	Inpatient hospital settings	59 (48.8)
	Emergency departments	56 (46.3)
	Outpatient clinics	30 (24.8)
Type of outcomes	Diagnostic	62 (51.2)
	Prognostic	59 (48.8)
Physician prediction mode	Probabilistic [1]	73 (60.3)
	Deterministic [2]	46 (38.0)
	Both	2 (1.7)
Physicians' experience †	Fully-trained physicians only	45 (37.2)
	Physicians in training only	18 (14.9)
	Mixed level	39 (32.2)
	Not available	19 (15.7)
Physicians' specialty	Surgical specialties	20 (16.5)
	Medical specialties	68 (56.2)
	Surgical and medical specialties	2 (1.7)
	Not reported	31 (25.6)
Statistical models	Local [3]	34 (28.1)
	Imported	87 (71.9)
Accuracy measures †	AUC/ORC [4]	59 (48.7) [5]
	Overall accuracy [6]	86 (71.1) [7]
	Sensitivity and specificity	76 (62.8)
	Odds ratio	76 (62.8)
	Calibration measures [8]	17 (14.0)
Total		121 (100.0)

† Non-exclusive alternatives. [1] Predictions about the probability of outcomes via numerical estimates or classification into non-quantitative risk categories (e.g. "likely" or "unlikely"). [2] Predictions about the current or future occurrence of the outcomes in binary terms (e.g. "present" versus "absent"). [3] Statistical models derived for patients from the same centre where comparisons were conducted. [4]AUC: Area Under Received Operational Characteristics curve. ORC: Ordinal C-Index. [5] In 28 comparisons AUCs were the only measure available. [6] Overall Accuracy (True positive predictions + True Negative predictions / Total predictions). [7] In 14 comparisons OAs were the only measure available because predictive tasks consisted of deterministic predictions over sets of nominal categories (e.g. differential diagnoses). [8] Hosmer–Lemeshow goodness-of-fit test and standard calibration plots.

Figure 1: Flow diagram of study selection process



Adapted from PRISMA (Transparent report of systematic reviews and meta-analyses). [1] Studies where the information available was artificially restricted to pre-selected sets of relevant data (e.g. case summaries) were not included.[2]Access to structured data-collection forms was admissible only if this was not the only source of information and the estimates provided by statistical models were not made available to physicians by researchers.

6.4.2. Quality analysis

The quality of studies is summarized in table 2. Predictions were prospective in all except 17 comparisons (14.0%), in which statistical models were applied retrospectively. Enrolment was consecutive in 89 comparisons (73.5%). Reference standards were considered adequate for most comparisons (102, 84.3%). In 51 comparisons (42.1%) outcomes were confirmed using the same test or, in case of objective outcomes such as death, with complete follow-up. A composite of tests was used in 59 comparisons (48.7%), and in 11 comparisons (9.1%) confirmation was unclear or not attained. Most comparisons (102, 84.3%) were considered free of incorporation bias. In 62 comparisons (51.2%), a blind application of the reference standard could not be ascertained. Forty-three comparisons (35.5%) came from studies where missing data was less than 5%. Sample-size calculations were present in only 13 comparisons (10.7%).

6.4.3. Predictive accuracy

The predictive accuracy of physicians and models across different measures of discrimination is shown in table 3. Due to the absence of adequate tests, a substantial number of comparisons was considered inconclusive (AUCs/OCDs: 41.1%, OAs: 46.5%, odds ratios: 89.5%). Among the remaining 69 comparisons, 18 (26.1%) favoured models, 32 favoured physicians (46.4%) and 19 (27.5%) indicated no significant differences in terms of AUCs/OCDs or OAs. These proportions allow rejection of the hypothesis that most comparisons show statistical models to be significantly better than physicians' judgement (binomial exact test, p=0.000118). In contrast, these findings are consistent with the hypothesis that physicians discriminate better than models in most comparisons (binomial exact test, p=0.631). The sensitivity of statistical models was more often superior to that of physicians (12 comparisons favouring models versus 8 comparisons favouring physicians) but the specificity of physicians was more frequently superior to that of models (29 comparisons favouring physicians versus 14 comparisons favouring models).

The application of the inductive threshold over the total set of comparisons (n=121) yielded 56 comparisons (46.3%) suggesting no difference, 43 (35.5%) indicating that physicians discriminated better than models, and 22 (18.2%) favouring models' discrimination in terms of AUCs or OAs (Table 3), again suggesting the superiority of physicians over statistical models. When the sample was reduced to 95 non-redundant predictive tasks via random-effects models applied to homogeneous tasks, analyses in terms of AUCs or OAs using both standard and inductive thresholds yielded similar results (Appendix 2). Likewise, when overrepresented outcomes were excluded from the analysis the data yielded analogous findings (Appendix 2).

6.4.4. Subgroup analyses

Pre-specified subgroup analyses applying the inductive threshold to AUC/OCDs or OAs indicated that imported statistical models outperformed physicians even more rarely (in only 12 (13.8%) of the 87 comparisons) but this inferiority was reversed in the smaller subgroup of local models (n=34 comparisons), in which 10 (29.4%) comparisons favoured models versus 6 (17.6%) comparisons favouring physicians. This reduction in the proportion of comparisons favouring physicians in the subset of local models was statistically significant (Fisher's exact test, p=0.011) (Table 4). The Cochran–Armitage test for trend suggested that the proportion of comparisons in which models were significantly better than physicians decreased with increasing level of training (p=0.024) (Table 5).

6.4.5. Calibration analysis

Of 38 studies investigating probabilistic predictions, only 12 compared physicians' calibration with that of models. In 9 of the 17 comparisons reported (52.9%) models were better calibrated than physicians, who tended in most cases to overestimate outcomes. Models' superior calibration, however, was accompanied by better discrimination in only 3 of 9 comparisons.

Table 2: Study quality

Author (year)	N Comparisons included	Study design	Type of enrolment	Sample Size calculation	Reference Standard	Risk of incorporation bias	Outcome confirmation all sample	Blinding †	Precision measures ‡	Statistical Test(s)	Missing data <5% ††
De Groot et al [2014]	4	Prospective	C	Not stated	Adequate	No	Yes[4]	No or unclear	Yes	No	Yes
Easter et al [2014]	3	Prospective	U	Yes	Adequate	No	Yes[4]	No or unclear	Yes	No	No
Jain et al [2014]	3	Prospective	C	Not stated	Adequate	No	Yes[5]	Yes[7]	No	Yes[8]	Yes
Mán et al [2014]	1	Prospective	Unclear	Not stated	Adequate	No	No	Unclear	Yes	Yes[11]	Unclear
Thompson et al [2014]	5	Prospective	C	Not stated	Adequate	No	Yes	Yes	Yes	No	No
Chew et al [2013]	7	Prospective	Not C	Not stated	Adequate	No	Yes[4]	Yes	No	Yes[8]	Yes
Farion et al [2013]	2	Prospective	Not C	Not stated	Adequate	No	Yes	No or unclear	Yes	Yes[9]	No
Laurent et al [2013]	6	Prospective	C	Not stated	Adequate	No	Yes[5]	Yes[7]	Yes	No	Yes
Meltzer et al [2013]	1	Prospective	Not C	Not stated	Unclear	No	Yes[4]	Unclear	Yes	No	No
Peñaloza et al [2013]	2	Mixed[1]	C	Not stated	Adequate	No	Yes[4]	No or unclear	Yes	Yes[11]	No
Wang et al [2013]	1	Prospective	Unclear	Not stated	Adequate	No	Yes[4]	Yes	Yes	No	Unclear
Litton et al [2012]	2	Prospective	C	Not stated	Adequate	No	Yes[5]	Yes[7]	Yes	No	No
Peñaloza et al [2012]	1	Mixed[1]	Not C	Not stated	Unclear	No	Yes[4]	No or unclear	Yes	No	No
Bruins Slot et al [2011]	1	Prospective	C	Not stated	Adequate	No	Yes[6]	No or unclear	Yes	No	Yes
Dionne et al [2011]	2	Prospective	C	Not stated	Adequate	No	Yes[6]	No or unclear	Yes	Yes[8]	No
Tenorio et al [2011]	1	Mixed[1]	Unclear	Not stated	Adequate	No	Unclear	No or unclear	Yes	No	Yes
Geersing et al [2010]	1	Mixed[1]	C	Not stated	Adequate	No	Yes[4]	Yes	Yes	No	Yes
Lintula et al [2010]	2	Prospective	C	Yes	Adequate	No	Yes[4]	No or unclear	Yes	Yes[10]	Yes
Chan et al [2009]	1	Prospective	C	Not stated	Adequate	No	Yes[4]	No or unclear	Yes	No	Yes
Kabrhel et al [2009]	1	Prospective	Mixed[2]	Not stated	Unclear	No	Yes[4]	Yes	Yes	No	Yes

Lintula et al [2009]	2	Prospective	C	Not stated	Adequate	No	Yes[4]	No or unclear	Yes	Yes[10]	Yes
Kline et al [2008]	1	Prospective	Mixed[2]	Yes	Unclear	No	Yes[4]	No or unclear	Yes	No	No
Cooper et al [2007]	1	Prospective	Unclear	Not stated	Unclear	No	Yes[4]	No or unclear	Yes	No	Yes
Van Gerven et al [2007]	3	Prospective	Unclear	Not stated	Unclear	Unclear	Unclear	No or unclear	Yes	Yes[11]	Unclear
Carrier et al [2006]	2	Mixed[1]	C	Not stated	Unclear	Yes	Yes[4]	Yes	Yes	No	Unclear
Hadjianastassiou et al [2006]	2	Prospective	C	Not stated	Adequate	No	Yes[5]	Yes[7]	Yes	Yes[11]	Yes
Mitchell et al [2006]	2	Prospective	C	Yes	Adequate	No	Yes[4]	No or unclear	Yes	No	No
Kabrhel et al [2005]	1	Prospective	Unclear	Not stated	Adequate	No	Yes[4]	Yes	Yes	No	Yes
Quinn et al [2005]	1	Prospective	Unclear	Not stated	Adequate	No	Yes[4]	No or unclear	Yes	No	Yes
Stein et al [2005]	1	Prospective	C	Not stated	Adequate	No	Yes[6]	Yes	Yes	Yes[9]	No
Blattler et al [2004]	1	Mixed[1]	C	Not stated	Adequate	No	Yes[4]	No or unclear	Yes	No	Yes
Pruekprasert et al [2004]	1	Prospective	Unclear	Not stated	Adequate	No	No	No or unclear	Yes	No	Unclear
Scholz et al [2004]	2	Prospective	C	Not stated	Adequate	No	Yes[5]	Yes[7]	Yes	No	Yes
Al Omar and Baldwin [2002]	1	Prospective	Unclear	Not stated	Adequate	No	Yes[6]	Yes	Yes	No	Unclear
Cornuz et al [2002]	1	Prospective	C	Not stated	Adequate	No	Yes[4]	Yes	Yes	No	No
Glas et al [2002]	2	Prospective	C	Yes	Adequate	No	Yes[6]	Yes	Yes	Yes[8]	Yes
Attia et al [2001]	1	Prospective	Unclear	Yes	Adequate	No	Yes[6]	Unclear	Yes	No	No
Bigaroni et al [2000]	2	Prospective	C	Not stated	Adequate	No	Unclear	Unclear	Yes	No	Unclear
Bojan et al [2000]	2	Prospective	Unclear	Not stated	Adequate	No	Yes[6]	Unclear	Yes	No	Unclear
Marcin et al [2000]	3	Prospective	C	Yes	Adequate	No	Yes[5]	Yes[7]	No	Yes[8]	Yes
Sanson et al [2000]	2	Prospective	C	Not stated	Adequate	No	Yes[4]	No or unclear	Yes	No	No
El Solh et al [1999]	1	Prospective	C	Not stated	Adequate	No	Yes[6]	No or unclear	Yes	No	Unclear
Pons et al [1999]	1	Prospective	C	Not stated	Adequate	No	Yes[5]	Yes[7]	No	Yes[11]	No
Beyth et al [1998]	1	Prospective	C	Not stated	Adequate	No	Yes[4]	Yes	Yes	No	No

Hallan et al [1997]	1	Prospective	C	Not stated	Adequate	No	Yes[4]	No or unclear	Yes	No	No
Richman et al [1997]	2	Prospective	C	Not stated	Adequate	No	Yes[4]	Yes	Yes	No	No
Brillman et al [1996]	1	Prospective	C	Not stated	Adequate	No	Yes[6]	Yes	Yes	No	Yes
Stevens et al [1994]	1	Prospective	C	Not stated	Adequate	No	Yes[5]	Yes[7]	No	Yes[8]	No
Detrano et al [1992]	2	Prospective	Unclear	Not stated	Adequate	No	Yes[6]	Unclear	No	Yes[8]	Unclear
Meyer et al [1992]	1	Prospective	Unclear	Not stated	Adequate	No	Yes[5]	Yes[7]	Yes	No	Yes
Baxt [1991]	1	Prospective	C	Yes	Adequate	No	Yes[4]	Yes	Yes	Yes[9]	No
Emerman et al [1991]	4	Mixed[1]	C	Not stated	Adequate	No	Yes[5]	Yes[7]	Yes	Yes[9]	Yes
Leivovici et al [1991]	1	Prospective	C	Not stated	Adequate	No	Yes[6]	Unclear	Yes	No	Yes
Bankowitz et al [1989]	1	Prospective	Unclear	Not stated	Unclear	Yes	Yes[4]	No or unclear	Yes	No	No
Brannen et al [1989]	1	Prospective	C	Yes	Adequate	No	Yes[5]	Yes[7]	Yes	Yes[8]	Yes
Chang et al [1989]	1	Prospective	C	Not stated	Adequate	No	Yes[5]	Yes[7]	Yes	No	No
Katzman-McClish et al [1989]	1	Prospective	Mixed[2]	Not stated	Adequate	No	Yes[5]	Yes[7]	Yes	Yes[8]	Yes
Sutton [1989]	3	Mixed [1]	Mixed[3]	Not stated	Unclear	Yes	Yes[4]	No or unclear	Yes	Yes[9]	No
Kruse et al [1988]	3	Prospective	C	Not stated	Adequate	No	Yes[5]	Yes[7]	Yes	Yes[8]	Unclear
Kirkeby and Riso [1987]	1	Prospective	Unclear	Not stated	Unclear	Unclear	Yes[4]	No or unclear	Yes	No	Unclear
Poretzky et al [1985]	1	Prospective	Unclear	Not stated	Adequate	No	Unclear	No or unclear	Yes	Yes[11]	No
Ikonen et al [1983]	1	Prospective	Not C	Not stated	Unclear	Unclear	Yes[4]	No or unclear	Yes	No	Unclear
Evenson et al [1975]	2	Mixed [1]	C	Not stated	Adequate	Unclear	Yes[4]	No or unclear	Yes	No	No
Horrocks and De Dombal [1975]	1	Prospective	C	Not stated	Adequate	Unclear	Yes[4]	No or unclear	Yes	No	Unclear
De Dombal et al [1975]	2	Prospective	C	Not stated	Adequate	Unclear	Yes[4]	No or unclear	Yes	No	Unclear
Oddie et al [1974]	2	Prospective	C	Not stated	Unclear	Unclear	No	No or unclear	Yes	No	Unclear
De Dombal et al [1972, 1974]	2	Prospective	C	Not stated	Adequate	No	Yes[4]	No or unclear	Yes	No	Unclear

Reale [1968]	1	Prospective	Unclear	Not stated	Unclear	Unclear	Yes[4]	No or unclear	Yes	No	Unclear

C: Consecutive. † Ascertainment of the outcome blind to both clinical and statistical predictions. ‡ Confidence intervals or standard errors available or possible to calculate for at least 1 measure per comparison. †† Cut off suggested by Schaffer (1999). 1. Prospective clinical prediction and retrospective statistical prediction for at least one model. 2. Consecutive and random. 3. Consecutive and convenience. 4. Complete follow-up, but confirmation based on different tests/observations. 5. Complete follow-up for an objective outcome (e.g. death). 6. Complete confirmation with the same test. 7. Objective outcome. 8. Hanley and McNeil (1983) or De Long et al (1988). 9. McNemar's test of proportions. 10. Chi-squared test (Randomized study, two independent samples) 11. Test not fully specified or inadequate.

Table 3: Discriminative predictive performance across all measures

	AUC/ORC n = 61 n (%)	Overall Accuracy n = 86 n (%)	Odds ratio n = 76 n (%)	Sensitivity n = 76 n (%)	Specificity n = 76 n (%)	AUC/ORC or if absent Overall Accuracy n = 121 n (%)
Standard threshold †						
Physicians better than models	12 (19.7)	29 (33.7)	6 (7.9)	8 (10.5)	29 (38.2)	32 (26.4)
Models better than physicians	8 (13.1)	14 (16.3)	2 (2.6)	12 (15.8)	14 (18.4)	18 (14.9)
No significant difference	16 (26.2)	3 (3.5)	0 (0)	3 (3.9)	1 (1.3)	19 (15.7)
Inconclusive	25 (41.1)	40 (46.5)	68 (89.5)	53 (69.7)	32 (42.1)	52 (43.9)
Inductive threshold ‡						
Physicians better than models	15 (24.6)	35 (40.7)	20 (26.3)	11 (14.5)	33 (43.4)	43 (35.5)
Models better than physicians	10 (16.4)	16 (18.6)	17 (22.4)	17 (22.4)	18 (23.7)	22 (18.2)
No difference	30 (49.2)	35 (40.7)	32 (42.1)	48 (63.2)	25 (32.9)	56 (46.3)
Inconclusive	6 (9.8)	0 (0)	7 (9.2)	0 (0)	0 (0)	0 (0)

† Statistically significant differences and their absence indicated by proper statistical test at p value < 0.05. If adequate test not available differences inferred from lack of overlap in 95% confidence intervals. Comparisons without adequate test and with overlap in 95% confidence intervals were considered inconclusive. AUC: Area under Receive Operative Characteristics curve. ORC: Ordinal C – Index. ‡ Statistical differences and their absence indicated by proper statistical test at p value < 0.05. If adequate test not available differences inferred from lack of overlap in 95% confidence intervals. Comparisons without test were considered as suggesting difference when there was less than 25% overlap in 95% confidence intervals. Overlaps equal or greater than 25% were considered as suggesting no difference.

Table 4: Discriminative predictive performance, local versus imported models ‡

	AUC/ORC n (%)	Overall Accuracy n (%)	Odds ratio n (%)	Sensitivity n (%)	Specificity n (%)	AUC/ORC or if absent Overall Accuracy n (%)
Imported models (n = 87)						
Physicians better than models	13 (28.3)	30 (50.0)	19 (32.8)	9 (15.5)	27 (46.6)	37 (42.5)
Models better than physicians	6 (13.0)	8 (13.3)	8 (13.8)	12 (20.7)	12 (20.7)	12 (13.8)
No difference	23 (50.0)	22 (36.7)	25 (43.1)	37 (63.8)	19 (32.8)	38 (43.7)
Inconclusive	4 (8.7)	0 (0)	6 (10.3)	0 (0)	0 (0)	0 (0)
Total	46 (100)	60 (100)	58 (100)	58 (100)	58 (100)	87 (100)
Local models (n = 34)						
Physicians better than models	2 (13.3)	5 (19.2)	1 (5.6)	2 (11.1)	6 (33.3)	6 (17.6)
Models better than physicians	4 (26.7)	8 (30.8)	9 (50.0)	5 (27.8)	6 (33.3)	10 (29.4)
No difference	7 (46.7)	13 (50.0)	7 (38.9)	11 (61.1)	6 (33.3)	18 (52.9)
Inconclusive	2 (13.3)	0 (0)	1 (5.6)	0 (0)	0 (0)	0 (0)
Total	15 (100)	26 (100)	18 (100)	18 (100)	18 (100)	34 (100)

AUC: Area under Receive Operative Characteristics curve. ORC: Ordinal C – Index. ‡ Statistical differences and their absence indicated by proper statistical test at p value < 0.05. If adequate test not available differences inferred from lack of overlap in 95% confidence intervals. Comparisons without test were considered as suggesting difference when there was less than 25% overlap in 95% confidence intervals. Overlaps equal or greater than 25% were considered as suggesting no difference.

Table 5: Discriminative predictive performance by level of training ‡

	AUC/ORC n (%)	Overall Accuracy n (%)	Odds ratio n (%)	Sensitivity n (%)	Specificity n (%)	AUC/ORC or if absent Overall Accuracy n (%) ‡
Physicians in training only (n = 18)						
Physicians better than models	2 (18.2)	5 (38.5)	4 (44.4)	2 (22.2)	3 (33.3)	5 (27.8)
Models better than physicians	2 (18.2)	3 (23.1)	0 (0.0)	1 (11.1)	2 (22.2)	5 (27.8)
No difference	7 (63.6)	5 (38.5)	5 (55.6)	6 (66.7)	4 (44.4)	8 (44.4)
Inconclusive	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Total	11 (100)	13 (100)	9 (100)	9 (100)	9 (100)	18 (100)
Mixed training level (n = 28)						
Physicians better than models	7 (35.0)	10 (40.0)	5 (17.9)	3 (10.7)	9 (32.1)	15 (38.5)
Models better than physicians	4 (20.0)	4 (16.0)	6 (21.4)	4 (14.3)	6 (14.3)	7 (17.9)
No difference	9 (45.0)	11 (44.0)	12 (42.9)	21 (75)	13 (46.4)	17 (43.6)
Inconclusive	0 (0.0)	0 (0.0)	5 (17.9)	0 (0.0)	0 (0.0)	0 (0.0)
Total	20 (100)	25 (100)	28 (100)	28 (100)	28 (100)	28 (100)
Fully trained physicians only (n = 34)						
Physicians better than models	6 (25.0)	16 (47.1)	10 (37.0)	5 (18.5)	15 (55.6)	19 (42.2)
Models better than physicians	0 (0.0)	5 (14.7)	7 (25.9)	7 (25.9)	6 (22.2)	3 (6.7)
No difference	12 (50.0)	13 (38.2)	10 (37.0)	15 (55.6)	6 (22.2)	23 (51.1)
Inconclusive	6 (25.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Total	24 (100)	34 (100)	27 (100)	27 (100)	27 (100)	45 (100)

AUC: Area under Receive Operative Characteristics curve. ORC: Ordinal C – Index. ‡ Statistical differences and their absence indicated by proper statistical test at p value < 0.05. If adequate test not available, differences inferred from lack of overlap in 95% confidence intervals. Comparisons without test were considered as suggesting difference when there was less than 25% overlap in 95% confidence intervals. Overlaps equal or greater than 25% were considered as suggesting no difference. ‡

6.5. Discussion

The main goal of this systematic review was to provide relevant empirical evidence to address the objection that the application of the DA in the context of diagnosis and prognosis might decrease the quality of the care of individual patients. In particular, my aim was to provide a methodologically robust empirical answer to the question of the alleged predictive superiority of statistical models over clinical judgment (e.g. Grove et al., 2000; Howick, 2011).

Our systematic review did not corroborate previous findings suggesting that the predictive performance of statistical models is generally superior to physicians' judgment. Rather, it suggests that, by and large, physicians' predictions are more accurate than predictions made by statistical models. Analyses with both the standard and the inductive thresholds indicated that physicians were nearly twice as likely as models to show superior AUCs or OAs. Furthermore, this study revealed that the training level of physicians, and whether the model was used in its local or an external setting influenced the relative accuracy of predictions made by physicians and statistical models.

This systematic review was designed to provide a comprehensive update on the topic of prediction in diagnosis and prognosis. Our inclusion criteria were designed to address the limitations of previous reviews, and to maximize the generalizability and relevance of the findings to the medical field. Towards these ends, we made several pragmatic decisions that deserve discussion.

First, our choice of comparisons as the main unit of analysis was motivated by the presence of many sources of heterogeneity, including: patient and physician characteristics, predictive task circumstances, types of outcomes, reference standards, accuracy measures, and statistical analyses.[123] Second, in order to facilitate the interpretation of comparisons without statistical tests for paired samples we complemented standard inference with a relatively conservative inductive threshold (§ 6.3.4).[124] An alternative would have been to ignore inconclusive comparisons, or to have relied on tests for unpaired samples (Grove et al., 2000). We decided against these options to avoid eliminating valuable information while being explicit as to the adequacy of statistical inferences across primary comparisons.

[123] Because of the possibility that overrepresented outcomes (e.g. mortality) and models (e.g. Wells' score) might have biased the overall result, we carried out supplementary analyses. Both the analysis by non-repetitive predictive tasks (§ 6.7.2, Table 7) and the exclusion of overrepresented outcome groups (§ 6.7.2, Tables 8 and 9) indicated findings similar to the main analysis.

[124] Although the use "inductive thresholds" is not common in medical research (Siontis et al., 2012), in this case it was well justified because it made it possible to obtain trends for the whole dataset that were both plausible and consistent with findings in the subset of comparisons where standard statistical inferences were possible.

Third, variation in the design and reporting of primary studies meant that we could not apply one discrimination measure across the whole set of comparisons. We acknowledge that AUCs and OAs as primary measures have important limitations [125] (Alberg et al., 2004; Cook, 2008; Van Calster et al., 2012; Mallett et al., 2012), however, they were chosen so to maximize the use of the information available. The relative superiority of physicians over models was less clear in terms of odds ratios; this suggests that the advantage of clinical judgment derived in part from outcome prevalences. However, the absence of strong reasons to suspect unrepresentativeness (studies were done in routine practice and enrolment was typically consecutive) argues that OAs will be a good indication of overall discriminatory performance in real clinical settings.[126]

The quality of included studies also deserves comment. Although several studies had important methodological drawbacks such as incomplete reporting and small sample sizes, quality analyses did not suggest the presence of significant biases in any direction. This implies that our findings are not likely to be biased but inherent limitations related to statistical estimation inherited from primary studies introduce some uncertainty into our conclusions.[127]

Our findings are at odds with previous reviews (e.g. Grove et al., 2000; Ægisdóttir et al., 2006; Marchese, 1992), which have favoured statistical models over physicians' judgment and have been used by advocates of EBM (a) to defend the development and application of statistical models as one of the cornerstones of the EBM approach in the context of diagnosis and prognosis (e.g. Perel et al., 2013; Gottlieb, 2009; Adams and levenson, 2012; Chew, 2014), (b) to cast doubt on the clinical relevance of physician's judgment and derivatively of an approach such as the DA (Howick, 2011, Eddy, 1990a), and (c) to substantiate the rationale of EBM hierarchies that place statistical models over physicians' judgment (McGinn et al., 2008; Phillips et al., 2009).

[125] In particular, the practical import of AUCs can be difficult to interpret and AUCs may even be misleading if they depend substantially on clinically irrelevant cut-offs. OAs are also liable to arbitrariness, can easily conceal variations in the frequency of different types of classificatory errors, and depend heavily on sample prevalences.

[126] Notice that, in the subset of comparisons with paired measures of discrimination, models typically had better sensitivity and physicians better specificity. This may suggest that models are better at "ruling out" outcomes than physicians, and that physicians are better at "ruling in" outcomes than models. However, it should be noted that the relevant frequencies often depended on arbitrary cut-offs; thus, these trends can be easily inverted. In addition, in several studies in which physicians made probabilistic judgements cut-offs were imposed subsequently by researchers, so true positive and true negative rates may not accurately reflect physicians' sensitivities and specificities in these studies.

[127] Publication bias could not be assessed statistically due to the heterogeneity of studies, but it is reasonable to think that if present it may favour models, given that studies are often done to demonstrate the usefulness of a particular model for clinical practice.

We are inclined to attribute this to the comprehensive analytic approach we have adopted, which was able to successfully address the limitations of previous reviews. These limitations include (1) measurements of statistical prediction on derivation samples, which overestimate the performance of statistical models, (2) reliance on simplified case summaries, which artificially eliminate the PEI and decreases the external validity of the findings (3) inclusion of medical students and non-medical professionals, which, among others, raises doubts about the kind of judgment on which predictions are based, and (4) the study of outcomes not pertaining to clinical medicine, which decreases the relevance of the findings to the point in question, namely: whether the performance of models or judgment is better in clinical contexts. Thus, it is important to stress that this systematic review offers the most updated and unbiased data available on the relative predictive performance of statistical models and physicians' judgment.

Notice also that the trends found in our subgroup analyses suggest directions for further research, which would clarify the scope of the application of clinical judgment and derivatively the DA. The apparent superiority of local over imported models can be interpreted in different ways. One possibility is that this superiority might have been due to "over-optimism" in research into local models (Moons et al., 2004; Siontis et al., 2012). However, this suggestion is in tension with the fact that all comparisons were made with new patients. A more obvious explanation is that physicians' relative advantage over imported models derives from their experience with local patterns of disease. This hypothesis is supported by recent evidence suggesting that prognostic discrimination worsens when new models are transported to different geographical areas (Siontis et al., 2015). Adequate studies are needed to confirm and further investigate this hypothesis, and in particular to investigate to what extent physicians' sensitivity to local patterns of disease are constrained by clinical complexity and the difficulty of gaining predictive feedback (Kahneman and Klein, 2009).

Further research is also needed into the influence of physicians' levels of training on predictive ability, which is crucial to determine how exactly physicians are able to use their judgment effectively to address the PEI, and clarify the practical import of the characteristic of the DA that I denoted "evidential flexibility". As I pointed out in the first chapter, the DA allows the physician to use different evidential sources to address the PEI and estimate the right probabilities for each individual (§ 1.5.4). With regard to "clinical experience", one of the potential sources to inform physicians' judgment in the context of the DA, a recent meta-analysis including studies in psychology suggests that it has a small but positive impact on accuracy (Spengler and Pilipis, 2015). However, the scarce data focused on physicians indicates that the relation between "clinical experience" and "predictive accuracy" is not consistent (Marcin et al., 1999; Conway et al., 2014; de Melo et al., 2013; Kong et al., 1989)

and may vary across different tasks. This does not necessarily undermines "evidential flexibility" as a virtue in the context of the DA but it certainly calls for further study of the ways in which different kinds of clinical experience affect how physicians exercise their judgment to deal with the PEI.

Nevertheless, despite the abovementioned qualifications, a number of practical lessons are implied by this systematic review. First, a preference for standardized prediction tools over physicians' informal predictions based on belief that the former has superior predictive accuracy is not warranted by our findings. Such belief, which as I said, has been endorsed by supporters of EBM is likely to be based on misconceptions about clinical judgment and statistical models demands revision or at least qualification.

Second, the fact that physicians in local settings generally outperform statistical models suggests that in the context of diagnosis and prognosis the PEI acquires practical importance for the EBM approach. Of course, one could interpret the problems of statistical models as a matter of lack of external validity and calibration (Sampson et al., 2009), but this interpretation fails to consider the superior performance of physicians' in local settings could also be explained by their ability to estimate the right probabilities when they are familiar with the extra information that needs to be accounted for by exercising clinical judgment.

Third, this systematic review also has implications for our discussion on the relationship between EBGs and standards of care and quality measures (§ 2.4.2). For, according to our findings, judgements about particular cases that diverge from diagnostic or prognostic EBGs based on statistical models should not automatically be viewed as sub-optimal. This not only calls for new diagnostic and prognostic EBGs[128], which should now pay more attention to evidence about the comparative accuracy of physicians' judgments, but it also calls for a richer and more nuanced interpretation of the notion of "best" (in this case diagnostic and prognostic) practices, one that do not uses incentives to reward compliance with methods that are less accurate or that forces physicians to rely on methods of inferior accuracy in order to protect themselves from malpractice lawsuits.[129]

Fourth, although I take these findings to call into question an automatic reliance on predictive models and highlight the discriminative accuracy of clinical judgment, I do not take them to show that statistical models ought to be ignored. In fact, insofar as statistical models provide

[128] Current diagnostic and prognostic EBGs are typically designed to advise the reliance on statistical models, which is encouraged by placing these methods higher than clinical judgment in standard rankings of evidence (e.g. McGinn et al., 2008; Phillips et al., 2009).

[129] Notice that I do not take the results of this systematic review to show that statistical models are not useful. Even if they are generally less accurate than physicians, I recognise that there may be many reasons to rely on them, including efficiency.

the physician with information about the patient, the exercise of clinical judgment in accordance with the DA (that is, in the light of everything the physicians knows about the patient) demands attention to relevant information from models. As I illustrated in the first chapter with a clinical case (§ 1.5.1), the role of clinical judgment in diagnosis is to address the PEI, which does not necessarily imply ignoring information provided by statistical models.

Finally, with respect to a potential synergy between statistical models and clinical judgment, notice that the potential benefits of an approach such as the DA would also be vindicated in cases where physicians' judgment complements models' predictions. Nonetheless, since the prevailing idea endorsed by supporters of EBM was that physicians ought not to rely on their judgment because their judgment is biased and offers a worse predictive performance than models, I decided to focus on whether models generally discriminate better than physicians in real settings and leave investigation of synergies between physicians' judgment and statistical models as a further research project.

6.6. Conclusions

This chapter sheds light on the empirical worth of the DA in the context of diagnosis and prognosis. The systematic review addressed the question of whether statistical models outperform physicians with respect to predictive accuracy in real settings. Our data suggests that supporters of EBM are wrong to assume that the predictive accuracy of statistical models is typically superior to that of physicians. This suggests that the PEI constitutes a real problem for EBM in this context. The findings of this chapter thus offer an empirical vindication of the DA in the context of diagnosis and prognosis, and so complement the previous chapter's evidence that the DA would improve patient outcomes in the context of therapy.

6.7. Appendices

6.7.1. Search Strategy

1. Statistical prediction.mp. 2. mechanical prediction.mp. 3. actuarial prediction.mp. 4. algorithmic prediction.mp. 5. Computer-aided.mp. 6. exp Decision Support Techniques/7. 1 or 2 or 3 or 4 or 5 or 6 8. diagnosis.mp. 9. diagnoses.mp. 10. diagnostic.mp. 11. prognosis.mp. 12. prognostic.mp. 13. risk prediction.mp. 14. exp risk assessment/ 15. 8 or 9 or 10 or 11 or 12 or 13 or 14 16. physician.mp. 17. physicians.mp. 18. doctor.mp. 19. doctors.mp. 20. practitioner.mp. 21. practitioners.mp. 22. medic.mp. 23. medics.mp. 24. psychiatrist.mp. 25. psychiatrists.mp. 26. neurologist.mp. 27. neurologists.mp. 28. surgeon.mp. 29. surgeons.mp. 30. p?ediatrician.mp. 31. p?ediatricians.mp. 32. gynecologist.mp. 33. gynecologists.mp. 34. obstetrician.mp. 35. obstetricians.mp. 36. internist.mp. 37. internists.mp. 38. cardiologist.mp. 39. cardiologists.mp. 40. pathologist.mp. 41. pathologists.mp. 42. radiologist.mp. 43. radiologists.mp. 44. dermatologist.mp. 45. dermatologists.mp. 46. gastroenterologist.mp. 47. gastroenterologists.mp. 48. anesthesiologist.mp. 49. anesthesiologists.mp. 50. ophthalmologist.mp. 51. ophthalmologists.mp. 52. rheumatologist.mp. 53. rheumatologists.mp. 54. endocrinologist.mp. 55. endocrinologists.mp. 56. geriatrician.mp. 57. geriatricians.mp. 58. Otolaryngologist.mp. 59. Otolaryngologists.mp. 60. neonatologist.mp. 61. neonatologists.mp. 62. urologist.mp. 63. urologists.mp. 64. immunologist.mp. 65. immunologists.mp. 66. nephrologist.mp.67. nephrologists.mp.68. intensivist.mp.69. intensivists.mp. 70. clinical prediction.mp. 71. medical prediction.mp. 72. clinical judg?ment.mp. 73. exp Judgment/ 74. judgment.mp.

75. 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 or 34 or 35 or 36 or 37 or 38 or 39 or 40 or 41 or 42 or 43 or 44 or 45 or 46 or 47 or 48 or 49 or 50 or 51 or 52 or 53 or 54 or 55 or 56 or 57 or 58 or 59 or 60 or 61 or 62 or 63 or 64 or 65 or 66 or 67 or 68 or 69 or 70 or 71 or 72 or 73 or 74.

76. 7 and 15 and 75

MEDLINE (05/09/15): 2860

EMBASE (05/09/15): 3517

Psyco (05/09/15): 129

Total: 6506.

Total after duplicates were removed: 5008.

Total after additional records were added: 5069

6.7.2. Supplementary analyses

In order to establish whether overrepresented predictive tasks might have exerted a strong influence over the overall result, we pooled the outcomes obtained in sets of comparisons involving homogeneous predictive tasks; that is, comparisons involving the same type of population (e.g. adults), the same outcome (e.g. hospital mortality), the same setting (e.g. critical care units) and the same model (e.g. model based on APACHE II score). By pooling

the results of each uniform set, we summarized 37 "individual comparisons" into 11 "pooled comparisons" (Table 7). The total set of comparisons in this analysis was thus reduced from 128 to 95.

Table 7: 11 sets of homogeneous predictive tasks †

Predictive tasks (studies)	Input comparisons		Pooled accuracy measures				
			AUC	Sen	Spe	OA	OR
AA in adults, Alvarado Score [40, 45, 68]	3	Physicians	-	0.93 (0.73-0.98)	0.79 (0.57-0.92)	0.88 (0.68-0.96)	48.5 (65-359.5)
		Model	-	0.75 (0.69-0.81)	0.72 (0.47-0.88)	0.73 (0.56-0.86)	7.1 (2.7-18.6)
AA in adults, Lintula Score [54]	2	Physicians	-	0.94 (0.65-0.99)	0.81 (0.72-0.88)	0.87 (0.78-0.93)	79.1 (6.9-909.7)
		Model	-	0.86 (0.78-0.91)	0.85 (0.24-0.99)	0.84 (0.59-0.95)	44.5 (1.5-1349.2)
AA in children, Lintula Score [57]	2	Physicians	-	0.89 (0.72-0.96)	0.59 (0.44-0.72)	0.73 (0.58-0.84)	15.0 (1.8-124.0)
		Model	-	0.91 (0.53-0.99)	0.79 (0.55-0.92)	0.84 (0.58-0.95)	47.6 (1.5-1469.5)
PE in adults, Wells' Score [46,56,61,64,77]	5	Physicians	-	0.70 (0.45-0.87)	0.68 (0.52-0.80)	0.66 (0.59-0.72)	4.9 (3.0-8.2)
		Model	-	0.73 (0.64-0.80)	0.58 (0.43-0.72)	0.62 (0.52-0.72)	3.7 (2.1-6.5)
PE in adults, PERC Score [49,58]	2	Physicians	-	0.83 (0.54-0.95)	0.63 (0.48-0.75)	0.68 (0.65-0.71)	8.0 (3.5-18.2)
		Model	-	0.97 (0.92-0.99)	0.17 (0.07-0.37)	0.33 (0.27-0.40)	7.7 (5.3-11.3)
DVT in adults, Wells' Score [47,67,71,74]	6	Physicians	-	0.82 (0.54-0.95)	0.80 (0.62-0.91)	0.76 (0.63-0.86)	18.6 (8.6-40.4)
		Model	-	0.67 (0.54-0.78)	0.71 (0.57-0.81)	0.69 (0.61-0.76)	4.9 (3.6-6.6)
Mortality in ICU, APACHE II [48,91-93,95]	7	Physicians	-	0.65 (0.50 0.78)	0.87 (0.80 0.92)	0.79 (0.77-0.82)	12.5 (8.6 to 18.3)
		Model	-	0.61 (0.50 0.70)	0.89 (0.77 0.96)	0.79 (0.76-0.81)	13.0 (6.7 to 25.0)
Mortality in ICU, PRISM III-12 [76]	3	Physicians	0.92 (nr)	-	-	-	-
		Model	0.92 (nr)	-	-	-	-
Diagnosis in AAP, DIAG [94]	2	Physicians	-	-	-	0.66 (0.64-0.67)	-
		Model	-	-	-	0.59 (0.56-0.61)	-
Diagnosis in AAP, Leeds' system [96,98,103,104]	4	Physicians	-	-	-	0.74 (0.69-0.78)	-
		Model	-	-	-	0.80 (0.44-0.95)	-
Diagnosis in LGIS, Leeds' system [101]	2	Physicians	-	-	-	0.74 (0.53-0.88)	-
		Model	-	-	-	0.81 (0.72-0.87)	-

AA: Acute Appendicitis. PE: Pulmonary Embolism. DVT: Deep Venous Thrombosis. ICU: Intensive Care Unit. AAP: Patients with acute abdominal pain. LGIS: patients with lower gastro-intestinal symptoms. AUCs: Areas Under Received Operator Characteristics Curve. Sen: Sensitivity. Spe: Specificity. OA: Overall Accuracy. OR: Odds Ratio. nr: Not Reported. (-) Not available. † The pooled accuracies for the 11 predictive tasks sets were added to the remaining 84 single comparisons to form a total of 95 non-redundant comparisons. This latter dataset was analysed in terms of the number of comparisons favouring either physicians, models or suggesting no difference (table 8).

The strategy used to pool the results of different sets of predictive tasks was determined by the number of comparisons in need of pooling within each set and the accuracy measures available (Takwoingi et al., 2015). We pool sensitivities, specificities and odds ratios for sets with four or more comparisons using a bivariate random effects model (metandi command in Stata). In sets where only three or fewer comparisons were available we pooled overall accuracies, sensitivities and specificities using a univariate random effects model (Metaprop

command in Stata). Finally, we calculated pooled odds ratios for sets with less than four comparisons using a random effects model (Metan command in Stata).

In the set of 95 non-redundant comparisons, the standard threshold indicated that in terms of AUCs or, when absent, OAs, physicians were more often better discriminators than models (Table 8). The proportions obtained allow rejection of the hypothesis that most comparisons show statistical models to be significantly better than physicians' judgement (binomial exact test, p=0.0042). In contrast, these findings are consistent with the hypothesis that physicians discriminate better than models (binomial exact test, p=0.881). The inductive threshold failed to detect differences in most comparisons; but where differences were found, physicians tend to discriminate better than models (Table 8). These findings were consistent with the analysis using the whole set of comparisons.

Table 8: Analysis of 95 non-redundant predictive tasks

	AUC or if absent OA N (%)
Standard threshold †	
Physicians better than models	21 (22.1)
Models better than physicians	12 (12.6)
No significant difference	11 (11.6)
Inconclusive	51 (53.7)
Inductive threshold ‡	
Physicians better than models	26 (27.4)
Models better than physicians	15 (15.8)
No difference	50 (52.6)
Inconclusive result	4 (4.2)
Total	95 (100.0)

† Statistically significant differences and their absence indicated by proper statistical test at p value < 0.05. If adequate test not available differences inferred from lack of overlap in 95% confidence intervals. Comparisons without adequate test and with overlap in 95% confidence intervals were considered inconclusive. AUC: Area under ROC curve. OCD: Ordinal C – Index. ‡ Statistical differences and their absence indicated by proper statistical test at p value < 0.05. If adequate test not available differences inferred from lack of overlap in 95% confidence intervals. Comparisons without test were considered as suggesting difference when there was less than 25% overlap in 95% confidence intervals. Overlaps equal or greater than 25% were considered as suggesting no difference.

In addition, in order to assess whether certain outcomes that were more represented in primary studies might have biased the overall result we grouped comparisons in 24 sets of similar outcomes (Table 9). Unlike the analysis in terms of predictive tasks (table 7 and 8), in this case we were unable to obtain pooled estimates for every group of outcomes due to the presence of different accuracy measures (either AUCs or OA) across different studies. For this reason, we decided to test whether the exclusion of outcome groups with more than ten comparisons (Pulmonary Embolism and Deep venous thrombosis, mortality in intensive care units, surgical mortality, gastrointestinal diseases) would change the direction of findings.

Among the remaining 37 comparisons, 9 (%) favoured models, 17 favoured physicians (%) and 11 (%) indicated no significant differences in terms of AUCs/OCDs or OAs. These proportions obtained with the standard threshold, like those obtained in the main analysis (Table 4) and the analysis in terms of non-redundant predictive tasks (Table 8), allow rejection of the hypothesis that most comparisons show statistical models to be significantly better than physicians' judgement (binomial exact test, p=0.003) but are consistent with the hypothesis that physicians discriminate better than models (binomial exact test, p=0.741). The application of the inductive threshold typically failed to identify differences in discrimination (32 of 66 comparisons). However, where discriminatory differences could be ascertained physicians performed better than models (24 of 66 versus 11 of 66 comparisons).

Table 9: Number of comparisons and discriminative performance across 24 groups of similar outcomes

Outcomes groups (Studies)	N comparisons *	Standard threshold †	Inductive threshold ‡
GI bleeding intervention [37]	2	2 I	2 P
GI bleeding mortality [37]	2	2 I	2 P
Brain injury in children [38]	3	2 P, 1 I	2 P, 1 SM
Surgical mortality [39,44,62,79]	12	8 I, 3 SM, 1 ND	7 ND, 5 SM
Acute Appendicitis [40,45,54,57,68,81]	8	3 ND, 3 P, 1 SM, 1 I	4 ND, 3 P, 1 SM
Stroke outcomes [41]	5	4 I, 1 P	4 ND, 1 P
ACS mortality [42]	5	3 SM, 1 P, 1 ND	3 SM, 1 P, 1 ND
Bleeding [42,80]	3	2 ND, 1 SM	2 ND, 1 SM
Asthma exacerbation [43]	2	2 I	2 ND
PE or DVT in adults [46,47,49,53,55,56,58,61,64,67,71,74,77]	17	7 P, 7 I, 3 SM	9 P, 5 ND, 3 SM
ICU mortality [48,69,76,84,86,91-93,95]	15	7 ND, 5 P, 3 I	7 P, 8 ND
Coronary disease diagnosis [50,63,85,87,97]	7	3 I, 2 ND, 1 P, 1 SM	3 P, 3 ND, 1 SM
Back pain [51]	2	1 ND, 1 SM	1 ND, 1 SM
Gastro-Intestinal diseases [52,94,96,98,100,101,103,104]	11	5 I, 3 P, 3 SM	4 ND, 4 P, 3 SM
Self harm [59]	1	1 P	1 P
Carcinoid heart disease [60]	3	3 I	3 ND
Syncope outcome [65]	1	1 I	1 ND
Infectious diseases [66,73,75,78,89]	6	3 I, 2 P, 1 SM	3 ND, 2 P, 1 SM
Extremities fractures [70,72,82]	5	4 P, 1 ND	4 P, 1 ND
Admission and hospitalization [83,99]	3	1 I, 1 P, 1 SM	1 ND, 1 P, 1 SM
Pneumonia [88]	4	4 I, 1 SM	2 ND, 2 SM
Thyroid diagnosis [102]	2	2 I	2 ND
Congenital cardiac diagnosis [105]	1	1 I	1 ND
General differential diagnosis [90]	1	1 I	1 ND

Abbreviations: GI: Gastrointestinal. ACS: Acute Coronary Syndrome. PE: Pulmonary Embolism. DVT: Deep Venous Thrombosis. I: Inconclusive comparison. P: Physicians better than models. SM: Statistical models better than physicians. ND: No difference. Notes: * Discriminative accuracy measured in terms of Area under Received operating characteristic curves, Ordinal C – Indices or when absent Overall Accuracies. † Statistically significant differences and their absence indicated by proper statistical test at p value < 0.05. If adequate test not available differences inferred from lack of overlap in 95% confidence intervals. Comparisons without adequate test and with overlap in 95% confidence intervals were considered inconclusive. ‡ Statistical differences and their absence indicated by proper statistical test at p value < 0.05. If adequate test not available differences inferred from lack of overlap in 95% confidence intervals. Comparisons without test were considered as suggesting difference when there was less than 25% overlap in 95% confidence intervals. Overlaps equal or greater than 25% were considered as suggesting no difference.

6.7.3. Predictive discriminative performance by study

Study (year)	Predictive task	Outcome	Confirmation	Approach	Predictors	% Prevalence (Events /Patients)	AUC (95% CI)	Cut-off	Sensitivity (95% CI)	Specificity (95% CI)	Overall Accuracy (95% CI)	Odds Ratio (95% CI)
De Groot et al (2014)	Before endoscopy, probability of GI bleeding requiring intervention.	Any intervention to control bleeding or mortality.	Follow-up. Bleeding established by composite of Sx/lab/endoscopy.	Clinical [a]	Consultant gastroenterologists	68.8 (667/970)	0.77 (NR)	≥ 1%	0.83 (0.79-0.85)	0.60 (0.54-0.65)	0.75 (0.73-0.78)	7.0 (5.2-9.6)
				Statistical	Blatchford Score	68.8 (667/970)	0.86 (NR)	≥ 1	0.99 (0.98-1.00)	0.14 (0.10-0.18)	0.73 (0.70-0.75)	25.9 (9.2-73.1)
	Before endoscopy, probability of mortality.	All cause mortality. (30 days).	Follow-up.	Clinical [a]	Consultant gastroenterologists	4.5 (44/968)	0.68 (NR)	≥ 1%	0.77 (0.62-0.88)	0.53 (0.50-0.56)	0.54 (0.51-0.57)	3.8 (1.9-7.8)
				Statistical	Blatchford score.	4.6 (43/943)	0.67 (NR)	> 2	0.95 (0.83-0.99)	0.13 (0.11-0.16)	0.17 (0.15-0.20)	3.2 (0.8-13.3)
	After endoscopy, probability of GI Bleeding requiring intervention.	Any intervention to control bleeding or mortality.	Follow-up. Bleeding established by composite of Sx/lab/endoscopy.	Clinical [a]	Consultant gastroenterologists	14.9 (140/941)	0.68 (NR)	≥ 1%	0.71 (0.62-0.78)	0.60 (0.56-0.63)	0.61 (0.58-0.65)	3.6 (2.4-5.3)
				Statistical	Rockall Score.	14.6 (140/956)	0.69 (NR)	> 2	0.95 (0.90-0.98)	0.23 (0.20-0.26)	0.34 (0.31-0.37)	5.8 (2.7-12.5)
	After endoscopy, probability of mortality.	All cause mortality (30 days).	Follow-up.	Clinical [a]	Consultant gastroenterologists	4.4 (41/937)	0.66 (NR)	≥ 1%	0.61 (0.45-0.75)	0.68 (0.65-0.71)	0.68 (0.65-0.71)	3.3 (1.8-6.4)
				Statistical	Rockall Score	4.6 (44/952)	0.76 (NR)	> 2	0.98 (0.86-1.00)	0.22 (0.19-0.24)	0.25 (0.22-0.28)	11.8 (1.6-86.5)
Easter et al (2014)	Brain injury in children with minor head trauma.	Composite (Including death, neurosurgery, and hospital admission).	CT or follow-up (Records or telephone interview).	Clinical [a]	Paediatric and EM staff	2.1 (21/1009)	0.94 (0.89-0.98)	≥ 1%	0.95 (0.76-1.00)	0.68 (0.65-0.71)	0.68 (0.66-0.71)	42.3 (5.7-316.9)
				Statistical	PECARN rule	2.1 (21/981)	0.81 (0.80-0.83)	+	1.00 (0.84-1.00)	0.62 (0.59-0.66)	0.63 (0.60-0.66)	71.3 (4.3-1180.8)
					CATCH rule	2.1 (21/1002)	0.67 (0.61-0.74)	+	0.91 (0.70-0.99)	0.44 (0.41-0.47)	0.45 (0.42-0.48)	7.4 (1.7-32.1)
					CHALICE rule	2.2 (19/858)	0.84 (0.76-0.93)	+	0.84 (0.60-0.97)	0.85 (0.82-0.87)	0.85 (0.82-0.87)	29.6 (8.5-103.1)
Jain et al (2014)	Probability of operative mortality after cardiac surgery.	Any death cause, or after 1 month, directly attributable.	Follow-up (death).	Clinical [a]	Cardiac surgeons	3.3 (168/5099)	0.73 (0.69-0.77)	-	-	-	-	-
				Statistical	Veterans Affairs Risk Score.	3.3 (168/5099)	0.78 (0.75-0.82)	-	-	-	-	-
	Probability of mortality after cardiac surgery.	Any death cause (1 year).	Follow-up (death).	Clinical [a]	Cardiac surgeons.	7.1 (360/5099)	0.61 (NR)	-	-	-	-	-
				Statistical	Veterans Affairs Risk Score.	7.1 (360/5099)	0.72 (NR)	-	-	-	-	-
	Probability of mortality after cardiac surgery.	Any death cause (5 years).	Follow-up (death).	Clinical [a]	Cardiac surgeons.	18.5 (942/5099)	0.64 (NR)	-	-	-	-	-
				Statistical	Veterans Affairs Risk Score.	18.5 (942/5099)	0.72 (NR)	-	-	-	-	-
Mán et al (2014)	Diagnosis of AA in adults.	Inflammation of the appendix.	Histologic confirmation and follow-up.	Clinical [b]	Staff surgeons	29.7 (41/138)	0.93 (NR)	BP	1.00 (0.89-1.00)	0.95 (0.88-0.98)	0.96 (0.91-0.99)	1395.9 (75.4 - 25834.9)
				Statistical	Alvarado score	25.2 (33/131)	0.75 (NR)	≥ 7	0.70 (0.51-0.84)	0.88 (0.79-0.93)	0.83 (0.75-0.89)	16.5 (6.3-42.9)
Thompson et al (2014)	Poor functional outcomes after ischemic stroke.	Oxford Handicap Scale score (6 months).	Follow-up (questionnaire by post).	Clinical [c]	18 staff and residents in geriatrics, IM and neurology.	35.2 (328/931)	0.74 (0.72-0.76)	≥ 3	0.44 (0.39-0.49)	0.96 (0.94-0.97)	-	18.9 (NR)
				Statistical	Reid's model	35.2 (328/931)	0.75 (0.73-0.77)	NR	0.45 (0.34-0.52)	0.96 (0.93-0.98)	-	19.6 (NR)
					Weimar's model	35.2 (328/931)	0.73 (0.71-0.76)	NR	0.43 (0.35-0.51)	0.96 (0.92-0.98)	-	18.1 (NR)
					SSV model	35.2 (328/931)	0.72 (0.70-0.74)	NR	0.43 (0.36-0.51)	0.95 (0.93-0.98)	-	14.3 (NR)
					Appelros' model	35.2 (328/931)	0.73 (0.71-0.75)	NR	0.42 (0.35-0.50)	0.95 (0.93-0.97	-	13.8 (NR)
					Lee's model	35.2 (328/931)	0.69 (0.66-0.71)	NR	0.38 (0.32-0.45)	0.94 (0.91-0.96)	-	9.6 (NR)
Chew et al	Probability of mortality	All cause mortality.	Follow-up	Clinical [a]	Physicians	3.1 (48/1542)	0.65 (0.60-0.71)	-	-	-	-	-

(2013)	after ACS (6 months).		(Records or phone).	Statistical	GRACE	3.1 (48/1542)	0.81 (0.77-0.85)	-	-	-	-	-
					PURSUIT	3.1 (48/1542)	0.43 (0.36-0.49)	-	-	-	-	-
	Probability of mortality after ST-MI (6 months).	All cause mortality.	Follow-up (Records or phone).	Clinical [a]	Physicians	Not stated	0.70 (NR)	-	-	-	-	-
				Statistical	TIMI	Not stated	0.78 (0.71-0.85)	-	-	-	-	-
	Probability of mortality after Non-ST-MI (6 months).	All cause mortality.	Follow-up (Records or phone).	Clinical [a]	Physicians	Not stated	0.66 (NR)	-	-	-	-	-
				Statistical	TIMI	Not stated	0.68 (0.60-0.75)	-	-	-	-	-
	Probability of mortality after ACS (1 month).	All cause mortality.	Follow-up (Records or phone).	Clinical [a]	Physicians	2.4 (37/1542)	0.71 (0.65-0.77)	-	-	-	-	-
				Statistical	GRACE	2.4 (37/1542)	0.79 (0.74-0.84)	-	-	-	-	-
	Probability of clinical bleeding event after ACS (6 months).	Composite (Including laboratory findings and interventions).	Follow-up (records or phone interview).	Clinical [a]	Physicians	4.7 (73/1542)	0.47 (0.43-0.52)	-	-	-	-	-
				Statistical	ACUITY	4.7 (73/1542)	0.51 (0.47-0.55)	-	-	-	-	-
					CRUSADE	4.7 (73/15f42)	0.51 (0.47-0.56)	-	-	-	-	-
Farion et al (2013)	Severity of exacerbation in children with asthma.	Mild (Stay less than 4 hrs.). Moderate (Stay 4–16 hrs.). Severe (Stay > 16 hrs. or admission).	Follow-up (Records and visit/phone contact after 10-14 days).	Clinical [c]	Staff, fellows and residents in EM.	68.3 (56/82)	-	Mod. +	0.79 (0.680.89)	0.77 (0.610.93)	0.78 (0.690.87)	12.2 (4.037.2)
				Statistical	PRAM	68.3 (56/82)	-	≥ 4	0.71 (0.600.83)	0.77 (0.610.93)	0.73 (0.640.83)	8.3 (2.824.6)
					Farion's model †	68.3 (56/82)	-	Mod. +	0.70 (0.580.82)	0.73 (0.560.90)	0.71 (0.610.81)	6.2 (2.217.6)
Laurent et al (2013)	Probability of operative mortality in adults with severe aortic stenosis surgery.	All cause mortality (30 days).	Follow up (Records, postal and telephone contact).	Clinical [a]	15 experienced cardiologists.	5.7 (18/314)	0.66 (0.53-0.80)	-	-	-	-	-
				Statistical	Additive EuroSCORE.	5.7 (18/314)	0.71 (0.59-0.83)	-	-	-	-	-
					Logistic EuroSCORE.	5.7 (18/314)	0.72 (0.60-0.84)	-	-	-	-	-
					EuroSCORE II	5.7 (18/314)	0.77 (0.65-0.89)	-	-	-	-	-
					STS Score	5.7 (18/314)	0.73 (0.61-0.86)	-	-	-	-	-
					Ambler Score	5.7 (18/314)	0.70 (0.59-0.81)	-	-	-	-	-
					ACEF score	5.7 (18/314)	0.66 (0.52-0.79)	-	-	-	-	-
Meltzer et al (2013)	Diagnosis of AA in adults.	AA defined radiologically or by surgical pathology.	Computer Tomography, laparotomy, follow-up.	Clinical [b]	Staff and residents in EM.	20.3 (53/261)	-	BP	0.79 (0.66-0.89)	0.68 (0.61-0.74)	0.70 (0.64-0.76)	8.0 (3.9-16.6)
				Statistical	Modified Alvarado score	20.3 (53/261)	0.69 (NR)	≥ 4	0.72 (0.57-0.83)	0.54 (0.47-0.61)	0.58 (0.52-0.64)	3.0 (1.6-5.8)
Peñaloza et al (2013)	Probability of PE in adults.	Composite including PE, VTE or related death.	Follow-up (3 months) Expert panel adjudication.	Clinical [a]	Staff and residents in EM.	31.3 (325/1038)	0.81 (0.78-0.84)	Mod. +	0.90 (0.86-0.93)	0.58 (0.54-0.61)	0.68 (0.65-0.70)	11.6 (7.9-17.1)
				Statistical	Wells Score	31.3 (325/1038)	0.71 (0.68-0.75)	≥ 2	0.81 (0.76-0.85)	0.60 (0.56-0.63)	0.66 (0.63-0.69)	6.4 (4.7-8.8)
					Revised GENEVA	31.3 (325/1038)	0.66 (0.63-0.70)	≥ 4	0.89 (0.85-0.92)	0.33 (0.30-0.37)	0.51 (0.47-0.54)	4.1 (2.8-6.0)
Wang et al (2013)	Probability of DVT in adults.	DVT defined by duplex scans and ultrasound.	Scans, ultrasound and follow up (45 days).	Clinical [a]	Vascular medicine specialists.	47.2 (191/405)	-	High risk	0.76 (0.70-0.82)	0.89 (0.84-0.93)	0.83 (0.79-0.86)	25.7 (15.0-44.1)
				Statistical	Wells Score	47.2 (191/405)	-	≥ 2	0.62 (0.55-0.69)	0.72 (0.65-0.78)	0.67 (0.63-0.72)	4.2 (2.8-6.4)
Litton et al (2012)	Mortality in critically ill patients.	Not stated.	Follow up.	Clinical [c]	12 Intensive care specialists	23.4 (585/2497)	-	< 2 yrs.	0.65 (0.61-0.69)	0.85 (0.84-0.87)	0.80 (0.79-0.82)	10.7 (8.7-13.2)
				Statistical	Knaus, model based on Apache II	23.4 (585/2497)	0.80 (0.78-0.82)	28%	0.65 (0.61-0.69)	0.81 (0.79-0.83)	0.77 (0.75-0.79)	7.8 (6.4-9.6)
					PREDICT	23.4 (585/2497)	0.80 (0.78-0.82)	20%	0.65 (0.61-0.69)	0.77 (0.75-0.79)	0.74 (0.72-0.76)	6.2 (5.1-7.6)
Peñaloza et al (2012)	Probability of PE in adults.	Composite including PE, VTE or related death.	Sequential testing (Roy et al. 2006).	Clinical [a]	Residents and specialists in EM.	29.8 (286/959)	-	Mod. +	0.91 (0.87-0.94)	0.55 (0.52-0.59)	0.66 (0.63-0.69)	12.4 (8.1-19.1)
				Statistical	PERC rule	29.8 (286/959)	-	≥ 1	8.2 (3.0-22.6)	0.37 (0.34-0.40)	0.99 (0.96-1.00)	0.10 (0.08-0.13)
Bruins Slot et al (2011)	Probability of ACS in adults.	ACS was defined in accordance with ESC and ACC guidelines.	Expert panel adjudication. Follow-up (1 month).	Clinical [a]	≈175 general practioners staff.	22.1 (66/298)	0.75 (0.68-0.82)	> 20%	0.85 (0.73-0.92)	0.34 (0.28-0.41)	0.45 (0.40-0.51)	2.9 (1.4-6.0)
				Statistical	Bruins Slot's model †	22.1 (66/298)	0.66 (0.58-0.73)	> 20%	0.74 (0.62-0.84)	0.51 (0.45-0.58)	0.56 (0.51-0.62)	3.0 (1.7-5.6)

Dionne et al (2011)	Probability of severe functional limitations (2 years) in patients with non-specific back pain.	Roland and Morris Disability questionary.	Follow up (telephone interview at 2 years).	Clinical [a]	Physicians	18.6 (195/1049)	0.69 (0.64-0.73)	≥ 50%	0.37 (0.31-0.45)	0.85 (0.83-0.88)	0.76 (0.74-0.79)	3.5 (2.4-4.9)
				Statistical	Cassandra rule (17 items)	18.6 (202/1085)	0.73 (0.69-0.77)	0.76	0.78 (0.72-0.84)	0.50 (0.47-0.54)	-	3.5 (NR)
					Cassandra rule (5 items) †	18.6 (201/1079)	0.78 (0.74-0.81)	0.80	0.79 (0.73-0.85)	0.60 (0.56-0.63)	-	5.6 (NR)
Tenorio et al (2011)	Diagnosis of celiac disease in children.	Not stated.	Specific history, serology, and histological findings.	Clinical [b]	≈ 18 trainees and staff in paediatric gastroenterology.	36.8 (14/38)	-	BP	0.64 (0.36-0.86)	0.96 (0.77-1.00)	0.84 (0.68-0.93)	41.4 (4.2-405.2)
				Statistical	Tenorio's model †	36.8 (14/38)	-	-	0.93 (0.64-1.00)	0.79 (0.57-0.92)	0.84 (0.68-0.93)	49.4 (5.2-473.4)
Geersing et al (2010)	Probability of DVT in adults.	Composite including PE, VTE or related death.	Sequential tests and follow-up (90 days). Expert panel adjudication.	Clinical [a]	≈ 300 General practitioners.	13.6 (136/1002)	0.82 (0.78-0.86)	> 20%	0.95 (0.89-0.98)	0.40 (0.37-0.43)	0.48 (0.44-0.51)	12.3 (5.7-26.7)
				Statistical	OUDEGA rule	13.6 (136/1002)	0.80 (0.77-0.84)	≥ 4	0.95 (0.89-0.98)	0.57 (0.54-0.60)	0.62 (0.59-0.65)	24.4 (11.3-52.7)
Lintula et al (2010)	Initial examination diagnosis of AA in adults.	AA defined by surgical and histological findings.	Surgery and histology or follow- up (1 month).	Clinical [b]	31 staff general surgeons.	44.4 (36/81)	-	BP	0.89 (0.73-0.96)	0.80 (0.65-0.90)	0.84 (0.74-0.91)	32.0 (9.0-114.0)
				Statistical	Lintula Score	54.2 (52/96)	-	≥ 21	0.87 (0.74-0.94)	0.59 (0.43-0.73)	0.74 (0.64-0.82)	9.3 (3.4-25.2)
	Final examination diagnosis of AA in adults.	AA defined by surgical and histological findings.	Surgery and histology or follow- up (1 month).	Clinical [b]	31 staff general surgeons.	44.4 (36/81)	-	BP	1.00 (0.88-1.00)	0.84 (0.70-0.93)	0.91 (0.82-0.96)	374.7 (20.7 - 6799.2)
				Statistical	Lintula Score	54.2 (52/96)	-	≥ 21	0.87 (0.74-0.94)	0.98 (0.86-1.00)	0.92 (0.84-0.96)	276.4 (32.6-2341.5)
Chan et al (2009)	Probability of DVT in pregnant women.	Composite: sequential testing and symptoms.	Ultrasound and follow up (3 months).	Clinical [a]	Thrombosis specialists.	8.9 (17/192)	-	Mod +	0.88 (0.62-0.98)	0.74 (0.66-0.80)	0.75 (0.68-0.81)	21.0 (4.6-95.5)
				Statistical	Chan (Based on LEFT rule) †	8.8 (17/194)	-	≥ 1	1.00 (0.81-1.00)	0.50 (0.43-0.58)	0.55 (0.47-0.62)	35.4 (2.1 to 597.7)
Kabrhel et al (2009)	Probability of PE in adults.	Composite including PE, VTE or attributable death	Ultrasound or V/Q scan. Angiography. Follow-up (45 days).	Clinical [a]	Physicians.	6.9 (545/7932)	-	Mod +	0.69 (0.65-0.73)	0.70 (0.69-0.71)	0.70 (0.69-0.71)	5.3 (4.4-6.5)
				Statistical	Wells score	6.9 (545/7940)	-	≥ 2	0.68 (0.64-0.72)	0.72 (0.71-0.73)	0.72 (0.71-0.73)	5.5 (4.5-6.6)
Lintula et al (2009)	Initial examination diagnosis of AA in children.	AA defined by surgical and histological findings.	Surgery and histology or follow- up (1 month).	Clinical [b]	31 staff general surgeons	45.0 (27/60)	-	BP	0.85 (0.65-0.95)	0.52 (0.34-0.69)	0.67 (0.53-0.78)	6.1 (1.7-21.6)
				Statistical	Lintula Score †	36.4 (24/66)	-	≥ 21	0.83 (0.62-0.95)	0.69 (0.53-0.82)	0.74 (0.62-0.84)	11.2 (3.2-39.2)
	Final examination diagnosis of AA in children.	AA defined by surgical and histological findings.	Surgery and histology or follow- up (1 month).	Clinical [b]	31 staff general surgeons	45.0 (27/60)	-	I	0.96 (0.79-1.00)	0.67 (0.48-0.81)	0.80 (0.67-0.89)	52.0 (6.2-435.1)
				Statistical	Lintula Score †	36.4 (24/66)	-	≥ 21	1.00 (0.83-1.00)	0.88 (0.74-0.96)	0.92 (0.82-0.97)	334.1 (17.7-6316.7)
Kline et al (2008)	Probability of PE in adults.	Composite including PE, VTE or related death.	Ultrasound or V/Q scan. Angiography/venography. Follow-up (45 days).	Clinical [a]	Physicians.	6.9 (561/8138)	-	Mod +	0.70 (0.66-0.74)	0.69 (0.68-0.70)	0.69 (0.68-0.70)	5.4 (4.5-6.5)
				Statistical	PERC rule	6.9 (561/8138)	-	≥ 1	0.96 (0.94-0.97)	0.25 (0.24-0.26)	0.30 (0.29-0.31)	7.6 (5.1-11.5)
Cooper et al (2007)	Probability of new episodes of self-harm (6 month).	Not stated.	Follow up (Records and databases).	Clinical [a]	Residents and staff in EM and psychiatry.	17.0 (1481/8722)	-	High risk	0.85 (0.83-0.87)	0.38 (0.37-0.39)	0.46 (0.45-0.47)	3.4 (2.9-4.0)
				Statistical	Manchester Self Harm Rule †	17.1 (1506/8825)	-	≥ 1	0.94 (0.92-0.95)	0.26 (0.24-0.27)	0.37 (0.36-0.38)	5.8 (4.6-7.3)
Van Gerven et al (2007)	Probability of heart disease in patients with low grade mid-gut carcinoid tumour.	Presence of moderate to extreme tricuspid valve insufficiency.	Follow up.	Clinical [a]	1 specialist in carcinoid tumours	40.7 (22/54)	0.66 (NR)	≥ 50%	-	-	0.69 (0.54-0.80)	-
				Statistical	Van Gerven (noisy threshold) †	40.7 (22/54)	0.66 (NR)	BP	-	-	0.72 (0.58-0.83)	-
					Van Gerven (naive bayesian) †	40.7 (22/54)	0.60 (NR)	BP	-	-	0.63 (0.49-0.75)	-

					Van Gerven (regression) †	40.7 (22/54)	0.59 (NR)	BP	-	-	0.67 (0.52-0.79)	-
Carrier et al (2006)	Probability of PE in adults.	PE or DVT.	V/Q scan, angiography, or CT. Ultrasonography. Follow-up (3 months).	Clinical [a]	Physicians	18.1 (78/432)	-	≥ 20%	0.86 (0.76-0.92)	0.38 (0.33-0.43)	0.47 (0.42-0.52)	3.8 (1.9-7.4)
				Statistical	Wells Score	18.4 (76/413)	-	> 4	0.83 (0.72-0.90)	0.41 (0.35-0.46)	0.48 (0.44-0.53)	3.3 (1.8-6.3)
					Rodger's model	19.8 (79/399)	-	≥ 1	0.96 (0.89-0.99)	0.29 (0.24-0.34)	0.42 (0.37-0.47)	10.2 (3.1-33.2)
Hadjianastassiou et al (2006)	Probability of postoperative mortality in adults with aortic aneurysm surgery.	Not stated.	Follow-up (death).	Clinical [a]	Residents in ICU	NR (NR/438)	0.82 (0.79-0.85)	-	-	-	-	-
				Statistical	Apache II model (LR) †	NR (NR/438)	0.87 (0.82-0.91)	-	-	-	-	-
					Apache II model (ANN) †	NR (NR/438)	0.87 (0.83-0.91)	-	-	-	-	-
Mitchell et al (2006)	Probability of acute coronary event in adults.	Coronary intervention, MI or death within 45 days.	Follow up (Records, telephone). Expert panel adjudication.	Clinical [a]	Staff and residents in EM	4.6 (51/1114)	0.78 (0.70-0.86)	> 2 %	0.96 (0.87-1.00)	0.27 (0.25-0.30)	0.31 (0.28-0.33)	9.2 (2.2-38.2)
				Statistical	ACI-TIPI.	4.6 (51/1114)	0.51 (0.44-0.58)	> 2 %	1.00 (0.93-1.00)	0.06 (0.05-0.08)	0.10 (0.08-0.12)	5.8 (0.4 - 94.8)
					Mitchell's model	4.6 (51/1114)	0.71 (0.65-0.78)	> 2 %	0.98 (0.90-1.00)	0.26 (0.24-0.29)	0.32 (0.29-0.35)	19.9 (2.7-145.0)
Kabrhel et al (2005)	Probability of PE in adults.	Radiological confirmation of PE.	Venous ultrasound, V/Q and CT scan, angiogram. Follow-up 3 months.	Clinical [a]	Staff and residents in EM, surgery and IM.	10.0 (61/607)	-	Most likely	0.54 (0.41-0.67)	0.76 (0.73-0.80)	0.74 (0.70-0.78)	3.8 (2.2-6.5)
				Statistical	Wells Score	10.0 (61/607)	-	> 4	0.59 (0.46-0.71)	0.78 (0.74-0.81)	0.76 (0.72-0.79)	5.0 (2.9-8.0)
Quinn et al (2005)	Probability of serious outcomes requiring admission (7 days).	Composite including death, AMI, PE, and stroke.	Follow up.	Clinical [a]	Residents and EM staff.	11.5 (79/684)	0.89 (0.85-0.93)	> 2%	0.94 (0.86-0.98)	0.52 (0.51-0.53)	0.57 (0.54-0.61)	16.4 (6.5-41.1)
				Statistical	San Francisco Rule †	11.5 (79/684)	0.92 (0.89-0.95)	≥ 1PA	0.96 (0.92-1)	0.62 (0.58-0.66)	0.66 (0.62-0.69)	41.3 (12.9-132.5)
Stein et al (2005)	Diagnosis of influenza infection in adults.	Influenza A or B presence.	Polymerase chain reaction test.	Clinical [b]	35 staff in EM and IM	24.4 (53/217)	-	BP	0.29 (0.18-0.43)	0.92 (0.87-0.95)	0.76 (0.70-0.82)	4.6 (2.0-10.4)
				Statistical	"Cough and Fever" rule	24.4 (53/217)	-	≥ 1 PA	0.40 (0.27-0.54)	0.92 (0.87-0.95)	0.79 (0.73-0.84)	7.6 (3.5-16.8)
Blattler et al (2004)	Probability of DVT in adults.	No flow detected or incompressible vein.	Ultrasound, venography. Follow-up (6 months).	Clinical [a]	Specialists in vascular medicine	27.7 (57/206)	-	High	0.81 (0.69-0.89)	0.85 (0.79-0.90)	0.84 (0.78-0.89)	24.1 (10.9-53.6)
				Statistical	Wells Score	27.7 (57/206)	-	≥ 1	0.54 (0.42-0.67)	0.84 (0.77-0.89)	0.76 (0.69-0.81)	6.2 (3.1-12.3)
Pruekprasert et al (2004)	Diagnosis of AA in adults.	AA defined by histological findings.	Surgery and histopathology.	Clinical [b]	Staff and residents in surgery	80.5 (186/231)	-		0.96 (0.91-0.98)	0.67 (0.50-0.80)	0.90 (0.85-0.93)	44.5 (17.4-114.1)
				Statistical	Alvarado Score	80.5 (186/231)	-	≥ 7	0.79 (0.72-0.85)	0.69 (0.53-0.81)	0.77 (0.71-0.82)	8.3 (4.0-17.2)
Scholz et al (2004)	Probability of survival in adults admitted to ICU.	Mortality during hospitalization.	Follow-up (death).	Clinical [a]	14 experts and fellows in ICM and IM residents.	17.7 (73/412)	0.84 (0.79-0.89)	-	-	-	-	-
				Statistical	SAPS II	17.7 (73/412)	0.75 (0.69-0.80)	-	-	-	-	-
					SAPS II (customized) †	17.7 (73/412)	0.72 (0.66-0.68)	-	-	-	-	-
Al Omar et al (2002)	Diagnosis of ankle or mid-foot fracture in children.	Any fracture needing medical intervention.	X - rays studies.	Clinical [b]	Physicians	21.3 (17/80)	-	BP	0.65 (0.38-0.86)	0.76 (0.63-0.86)	0.74 (0.63-0.83)	5.9 (1.9-18.6)
				Statistical	Ottawa ankle rule	21.3 (17/80)	-	≥ 1 PA	1.00 (0.81-1.00)	0.30 (0.19-0.43)	0.45 (0.34-0.56)	15.3 (0.9 - 268.1)
Cornuz et al (2002)	Probability of DVT in adults.	DVT (identified by ultrasound findings) and PE.	Ultrasound, D-dimer levels, others. Follow-up.	Clinical [a]	Residents in vascular medicine	29.5 (82/278)	0.72 (NR)	High	0.50 (0.39-0.61)	0.88 (0.82-0.92)	0.77 (0.71-0.81)	7.2 (3.9-13.2)
				Statistical	Wells Score	29.5 (82/278)	0.72 (NR)	≥ 1	0.83 (0.73-0.90)	0.48 (0.41-0.56)	0.59 (0.53-0.64)	4.6 (2.4-8.7)
Glas et al (2002)	Probability of ankle fracture in patients with acute ankle injury.	Any fracture diagnosed by the radiologist.	Radiographic series of foot and ankle.	Clinical [a, b]	General surgery residents	11.4 (74/647)	0.80 (0.74-0.87)	BP	0.82 (0.72-0.90)	0.68 (0.64-0.71)	0.69 (0.66-0.73)	9.8 (5.3-18.4)
				Statistical	Ottawa ankle rule	11.4 (74/647)	0.69 (0.62-0.76)	≥ 1 PA	0.89 (0.80-0.95)	0.26 (0.23-0.30)	0.33 (0.30-0.37)	2.9 (1.4-6.2)
					Leiden †	11.4 (74/647)	0.77 (0.71-0.83)	> 7	0.80 (0.69-0.88)	0.59 (0.55-0.63)	0.61 (0.57-0.65)	5.6 (3.1-10.1)
Attia et al (2001)	Probability of positive throat culture in children with acute pharyngitis.	Presence of Group A B-Haemolytic Streptococcus colonies at 24-48h.	Throat culture and serotyping.	Clinical [a]	Staff and residents in paediatric clinic and emergency service	37.1 (218/587)	-	> 5	0.72 (0.65-0.78)	0.60 (0.55-0.65)	0.65 (0.61-0.68)	3.9 (2.7-5.6)
				Statistical	Attia score †	38.5 (210/545)	-	≥ 4	0.18 (0.13-0.24)	0.97 (0.94-0.98)	0.66 (0.62-0.70)	7.0 (3.4-14.3)

Bigaroni et al (2000)	Probability of DVT in adults.	Not stated.	Dimer D, compression ultrasound and Lung scan.	Clinical [a]	Residents and specialists in vascular medicine	17.0 (28/165)	-	Mod +	1.00 (0.85-1.00)	0.46 (0.38-0.55)	0.55 (0.47-0.63)	48.6 (2.9 - 811.9)
				Statistical	Wells Score applied by residents	17.0 (28/165)	-	> 0	0.71 (0.51-0.86)	0.75 (0.67-0.82)	0.75 (0.67-0.81)	7.6 (3.1-18.8)
					Wells Score applied by specialists	17.0 (28/165)	-	> 0	0.77 (0.56-0.90)	0.73 (0.64-0.80)	0.73 (0.66-0.80)	8.9 (3.3-23.7)
Bojang et al (2000)	Diagnosis of malaria in febrile children.	Fever (≥37.5 °C) and a parasite count of ≥5000/ml.	Blood Film (Giemsa Stain).	Clinical [b]	1 experienced paediatrician	34.8 (133/382)	-	BP	0.82 (0.74-0.88)	0.61 (0.55-0.67)	0.68 (0.63-0.73)	7.1 (4.3-11.9)
				Statistical	Olaleye (Complete score)†	34.8 (133/382)	-	≥ 7	0.89 (0.83-0.94)	0.63 (0.57-0.69)	0.72 (0.67-0.77	14.5 (7.9-26.7)
					Olaleye (Symptoms) †	34.8 (133/382)	-	≥ 5	0.73 (0.64-0.80)	0.59 (0.53-0.65)	0.64 (0.59-0.69)	3.9 (2.5-6.1)
Marcin et al (2000)	Probability of mortality in children admitted to PICU.	Mortality during hospitalization.	Follow-up.	Clinical [a]	34 residents in paediatrics	6.7 (36/540)	0.92 (NR)	-	-	-	-	-
					9 Paediatric ICM fellows	6.0 (36/602)	0.87 (NR)	-	-	-	-	-
					5 Paediatric ICM staff	5.9 (36/612)	0.95 (NR)	-	-	-	-	-
				Statistical	PRMS III	5.7 (36/635)	0.92 (NR)	-	-	-	-	-
Sanson et al (2000)	Probability of PE in adults.	High-probability V/Q scan or an abnormal angiography.	V/Q scan, CT, angiography.	Clinical [a]	Staff physicians	30.5 (126/413)	-	> 80%	0.28 (0.20-0.37)	0.85 (0.81-0.89)	0.68 (0.63-0.72)	2.2 (1.3-3.7)
				Statistical	Wells score	29.5 (122/414)	-	≥ 2	0.66 (0.57-0.65)	0.36 (0.31-0.42)	0.45 (0.40-0.50)	1.1 (0.7-1.8)
					Wells score (extended)	37.6 (89/237)	-	Mod +	0.81 (0.71-0.88)	0.29 (0.22-0.37)	0.49 (0.42-0.55)	1.7 (0.9-3.3)
El-Solh et al (1999)	Diagnosis of pulmonary tuberculosis in adults.	Presence of M tuberculosis on respiratory specimens.	Auramine-rhodamine9 fluorescent stain and nuclei acid probes.	Clinical [b]	EM staff, infectious diseases fellows and medical residents.	9.2 (11/119)	0.72 (0.65-0.79)	BP	0.64 (0.31-0.89)	0.79 (0.72-0.87)	0.77 (0.69-0.84)	6.5 (1.7-24.0)
				Statistical	El-Solh's model †	9.2 (11/119)	0.92 (0.86-0.99)	BP	1.00 (0.72-1.00)	0.69 (0.61-0.78)	0.72 (0.63-0.80)	51.8 (3.0 - 905.6)
Pons et al (1999)	Probability of mortality in adults undergoing open-heart surgery.	Any death <30 days) or during the hospital stay.	Follow up (direct or telephone, and records).	Clinical [a]	Cardiac surgeons	NR (NR/359)	0.65 (NR)	-	-	-	-	-
				Statistical	Pons' model †	NR (NR/359)	0.70 (NR)	-	-	-	-	-
Beyth et al (1998)	Probability of major bleeding in adults treated with warfarin.	Overt bleeding requiring intervention.	Follow-up (various sources).	Clinical [a]	Physicians	10.2 (20/196)	-	Low +	0.55 (0.32-0.76)	0.45 (0.37-0.53)	0.46 (0.39-0.53)	1.0 (0.4-2.5)
				Statistical	Outpatient Bleeding Risk Index	10.2 (20/196)	-	High	0.25 (0.10-0.49)	0.95 (0.91-0.98)	0.88 (0.83-0.92)	7.0 (2.0-24.1)
Hallan et al (1997)	Probability of Acute Appendicitis in adults.	Histological examination.	Surgical and histological findings or follow up.	Clinical [a]	9 general surgery residents	36.5 (111/304)	0.81 (0.79-0.82)	-	-	-	-	-
				Statistical	Hallan's model. †	36.5 (111/304)	0.81 (0.80-0.83)	-	-	-	-	-
Richman et al (1997)	Probability of fracture in adults with knee injury.	Any fracture of the knee or patella seen on standard radiography.	Radiography and follow up (3 weeks).	Clinical [a]	Staff and residents in EM	7.4 (26/351)	-	≥ 20%	0.88 (0.69-0.97)	0.64 (0.58-0.69)	0.66 (0.60-0.70)	13.6 (4.0-46.3)
				Statistical	Bauer's rule	7.4 (26/351)	-	≥ 1 PA	0.85 (0.64-0.95)	0.49 (0.43-0.54)	0.52 (0.46-0.57)	5.3 (1.8-15.6)
					Stiell's rule	7.4 (26/351)	-	≥ 1 PA	0.85 (0.64-0.95)	0.50 (0.44-0.55)	0.52 (0.47-0.58)	5.5 (1.8-16.2)
Brillman et al (1996)	Probability of hospital admission in triage patients.	Actual rate of admission.	Admission determined by inpatient staff.	Clinical [a]	Staff and residents in EM	6.3 (242/3822)	-	Likely +	0.62 (0.55-0.68)	0.87 (0.86-0.88)	0.86 (0.84-0.87)	10.8 (8.2-14.3)
				Statistical	AMOS triage system	5.5 (195/3550)	-	Likely +	0.68 (0.61-0.75)	0.74 (0.72-0.75)	0.73 (0.72-0.75)	6.0 (4.4-8.1)
Stevens et al (1994)	Probability of mortality in neonates admitted to NICU.	Mortality during hospitalization in NICU.	Follow up.	Clinical [a]	18 NICU staff physicians.	4.0 (21/523)	0.85 (NR)	Mod +	0.90 (0.68-0.98)	0.68 (0.64-0.72)	0.69 (0.65-0.73)	20.1 (4.6-87.4)
				Statistical	SNAP-PE.	3.9 (21/524)	0.94 (NR)	-	-	-	-	-
Detrano et al (1992)	Probability of coronary disease in adults during angiography.	Any greater than 50% diameter obstruction.	Angiography (Expert panel adjudication).	Clinical [a]	Expert cardiologists	16.9 (67/397)	0.82 (NR)	-	-	-	-	-
				Statistical	Detrano's model. †		0.80 (NR)	-	-	-	-	-
	Probability of coronary disease in adults during angiography.	Triple vessel or left main obstructions.	Angiography (Expert panel adjudication).	Clinical [a]	Expert cardiologists	6.0 (24/397)	0.69 (NR)	-	-	-	-	-
				Statistical	Detrano's model. †	6.0 (24/397)	0.72 (NR)	-	-	-	-	-
Meyer et al	Mortality in adults admitted	Death prior discharge	Follow-up.	Clinical [b]	Intensive medicine staff	6.9 (40/578)	-	BP	0.55 (0.39-0.70)	0.98 (0.96-0.99)	0.95 (0.93-0.97)	64.5 (26.7-

(1992)	to SICU.	from SICU.										156.0)
				Statistical	Apache II score	6.9 (40/578)	-	≥ 20	0.58 (0.41-0.73)	0.94 (0.91-0.96)	0.91 (0.88-0.93)	20.1 (9.8-41.1)
Baxt et al (1991)	Diagnosis of MI in adults with anterior chest pain.	Composite.	Enzymes, ECG, scintiscan, and follow up (3 weeks).	Clinical [b]	EM staff and medical residents	10.9 (36/331)	-	BP	0.78 (0.60-0.89)	0.85 (0.80-0.89)	0.84 (0.80-0.88)	19.4 (8.3-45.4)
				Statistical	Baxt's model. †	10.9 (36/331)	-	> 0.55	0.97 (0.84-1.00)	0.96 (0.93-0.98)	0.96 (0.94-0.98)	903.6 (113.2-7211.8)
Emerman et al (1991)	Diagnosis of pneumonia in adults with respiratory infection symptoms.	Radiologic findings.	Postero-anterior and lateral chest radiographies (Interpreted independently by two radiologists).	Clinical [b]	39 staff physicians and residents.	7.2 (21/290)	-	BP	0.86 (0.64-0.97)	0.58 (0.51-0.64)	0.60 (0.54-0.65)	8.2 (2.3-28.4)
				Statistical	Gennis' rule	7.2 (21/290)	-	≥ 1 PA	0.62 (0.38-0.82)	0.76 (0.71-0.81)	0.75 (0.70-0.80)	5.2 (2.1-13.1)
					Diehr's model	7.2 (21/290)	-	> 0	0.67 (0.43-0.85)	0.67 (0.61-0.73)	0.67 (0.61-0.72)	4.0 (1.6-10.4)
					Heckering's model	7.2 (21/290)	-	> 1	0.71 (0.48-0.89)	0.67 (0.61-0.73)	0.68 (0.62-0.73)	5.1 (1.9-13.7)
					Singal's rule	7.2 (21/290)	-	> 0.26	0.76 (0.53-0.92)	0.55 (0.49-0.61)	0.57 (0.51-0.62)	3.9 (1.4-11.0)
Leibovici et al (1991)	Diagnosis of bacteraemia in febrile adults.	Positive (not contamined) blood culture.	Blood culture.	Clinical [b]	Internal medicine staff	14.2 (36/253)	-	BP	0.53 (0.36-0.69)	0.84 (0.79-0.89)	0.80 (0.74-0.85)	6.0 (2.8-12.7)
				Statistical	Leibovici's model †	14.2 (36/253)	-	≥ 20%	0.97 (0.84-1.00)	0.60 (0.53-0.66)	0.65 (0.59-0.71)	52.3 (7.0-388.9)
Bankowitz et al (1989)	Diagnosis in adults admitted to medical ward.	Various conditions.	Tests findings and subspecialist's opinion.	Clinical [d]	IM staff and residents	1.0 (20/20)	-	BP	-	-	0.30 (0.13-0.54)	-
				Statistical	Quick Medical Reference	1.0 (20/20)	-	BP	-	-	0.60 (0.36-0.80)	-
Brannen et al (1989)	Probability of mortality in adults admitted to ICU.	Mortality during hospitalization in ICU.	Follow up.	Clinical [a]	Fellows in PM and ICM	31.2 (34/109)	0.90 (NR)	> 50%	0.79 (0.62-0.91)	0.84 (0.73-0.91)	0.83 (0.74-0.89)	20.3 (7.2-57.0)
				Statistical	Knaus, model based on Apache II	31.2 (34/109)	0.80 (NR)	> 40 %	0.62 (0.44-0.77)	0.80 (0.69-0.88)	0.74 (0.65-0.82)	6.5 (2.6-15.8)
Chang et al (1989)	Mortality in adults admitted to ICU.	Mortality during hospitalization in ICU.	Follow-up.	Clinical [b]	2 specialists in ICM and surgery.	36.1 (82/227)	-	BP	0.36 (0.26-0.46)	0.96 (0.91-0.98)	0.73 (0.67-0.78)	13.8 (5.5-34.6)
				Statistical	Knaus, model based on Apache II	36.1 (82/227)	-	-	0.44 (0.34-0.55)	1.00 (0.97-1.00)	0.79 (0.73-0.84)	247.1 (14.9 - 4084.6)
Katzman et al (1989)	Probability of mortality in adults admitted to ICU.	Mortality during hospitalization in ICU.	Follow-up.	Clinical [a]	6 specialists in ICM	24.5 (128/523)	0.89 (NR)	> 50%	0.82 (0.74-0.88)	0.81 (0.76-0.84)	0.81 (0.77-0.84)	18.9 (11.3-31.6)
				Statistical	Knaus, model based on Apache II	24.5 (128/523)	0.83 (NR)	> 50%	0.55 (0.46-0.63)	0.92 (0.88-0.94)	0.83 (0.79-0.86)	13.2 (8.0-21.8)
Sutton et al (1989)	Diagnosis in adults with acute abdominal pain.	Various conditions.	Definite diagnosis assigned by consultant.	Clinical [d]	Junior doctors	100 (6379/6379)	-	BP	-	-	0.66 (0.64-0.67)	-
				Statistical	CAD-A		-	BP	-	-	0.57 (0.56-0.58)	-
					DIAG (retrospective)		-	BP	-	-	0.60 (0.58-0.61)	-
	Diagnosis in adults with acute abdominal pain.[1]	Various conditions.	Definite diagnosis assigned by consultant.	Clinical [d]	Junior doctors	100 (583/583)	-	BP	-	-	0.66 (0.61-0.69)	-
				Statistical	DIAG (prospective)		-	BP	-	-	0.57 (0.52-0.61)	-
Kruse et al (1988)	Probability of mortality in adults admitted to ICU.	Mortality during hospitalization in ICU.	Follow-up.	Clinical [a]	18 Medical interns	40 (146/366)	-	> 50%	0.64 (0.56-0.72)	0.90 (0.86-0.94)	0.80 (0.76-0.84)	17.1 (9.8-30.1)
					22 Residents in PM and ICM	40 (146/366)	-	> 50%	0.65 (0.57-0.73)	0.89 (0.84-0.92)	0.79 (0.75-0.83)	14.5 (8.5-24.9)
					17 Fellows in PM and ICM	40 (146/366)	0.89 (NR)	> 50%	0.65 (0.57-0.73)	0.89 (0.84-0.92)	0.79 (0.75-0.83)	14.5 (8.5-24.9)
				Statistical	Knaus, model based on Apache II	40 (146/366)	-	> 50%	0.67 (0.59-0.75)	0.86 (0.81-0.90)	0.79 (0.74-0.83)	12.9 (7.7-21.7)
Kirkeby et al (1987)	Diagnosis in adults with acute abdominal pain.	Various conditions.	Discharge diagnosis.	Clinical [d]	Physicians	100 (77/77)	-	BP	-	-	0.65 (0.53-0.75)	-
				Statistical	Leeds' decision support system	100 (77/77)	-	BP	-	-	0.53 (0.42-0.65)	-
Poretsky et al (1985)	Diagnosis of MI in adults with acute chest pain.	Composite including symptoms and test findings.	CK-MB, CK, SGOT, LDH, ECG, scintiscan.	Clinical [b]	Physicians	48.2 (81/168)	-	BP	0.74 (0.63-0.83)	0.85 (0.75-0.92)	0.80 (0.73-0.85)	16.3 (7.5-35.2)
				Statistical	Goldman's model	48.2 (81/168)	-	BP	0.81 (0.71-0.89)	0.53 (0.42-0.64)	0.67 (0.59-0.74)	4.9 (2.4-10.0)
Ikonen et al	Diagnosis in adults with	Various conditions.	Discharge diagnosis.	Clinical [d]	Surgeons	100 (290/290)	-	BP	-	-	0.76 (0.70-0.80)	-

(1983)	acute abdominal pain.			Statistical	Leeds' decision support system	100 (290/290)	-	BP	-	-	0.68 (0.63-0.74)	-
Evenson et al (1975)	Length of hospital stay in adults admitted to psychiatric hospital.	Short stay (< 90 days) and long stay (≥ 90 days).	Follow up (120 days).	Clinical [b]	Psychiatrists	41.1 (67/163)	-	≥ 90	0.72 (0.59-0.82)	0.79 (0.69-0.87)	0.76 (0.69-0.82)	9.6 (4.7-19.8)
				Statistical	Evenson's model †	41.1 (67/163)	-	≥ 90	0.64 (0.51-0.75)	0.75 (0.65-0.83)	0.71 (0.63-0.77)	5.4 (2.7-10.6)
	Unauthorized absence in adults admitted to psychiatric hospital.	Not stated.	Follow up (120 days).	Clinical [a]	Psychiatrists	11.4 (19/167)	-	≥ 5%	0.68 (0.44-0.86)	0.51 (0.43-0.60)	0.53 (0.45-0.61)	2.3 (0.8-6.3)
				Statistical	Evenson's model †	11.4 (19/167)	-	≥ 5%	0.37 (0.17-0.61)	0.75 (0.67-0.82)	0.71 (0.63-0.77)	1.8 (0.6-4.8)
Horrocks et al (1975)	Diagnosis in adults with upper-gastrointestinal symptoms.	Various conditions.	Final diagnosis (where relevant, histopathology).	Clinical [d]	Registrars and consultant surgeons	100 (122/122)	-	BP	-	-	0.93 (0.86-0.96)	-
				Statistical	Leeds' decision support system †	100 (122/122)	-	BP	-	-	0.88 (0.80-0.93)	-
De Dombal et al (1975)	Admission evaluation, diagnoses in adults with LAP.	Various conditions.	Final diagnosis (where relevant, histopathology).	Clinical [d]	Registrars and consultant surgeons	100 (301/301)	-	BP	-	-	0.64 (0.59-0.70)	-
				Statistical	Leeds' decision support system †	100 (301/301)	-	BP	-	-	0.77 (0.72-0.82)	-
	Perioperative evaluation, diagnoses in adults with LAP.	Various conditions.	Final diagnosis (where relevant, histopathology).	Clinical [d]	Registrars and consultant surgeons	100 (301/301)	-	BP	-	-	0.83 (0.78-0.87)	-
				Statistical	Leeds' decision support system †	100 (301/301)	-	BP	-	-	0.85 (0.80-0.88)	-
Oddie et al (1974)	Thyroid metabolic diagnosis In adults.	Hypothyroid, Euthyroid, Hyperthyroid.	Metabolic test data.	Clinical [c]	Physicians	100 (1066/1066)	-	BP	-	-	0.97 (0.96-0.98)	-
				Statistical	Oddie's model	100 (1066/1066)	-	BP	-	-	0.97 (0.96-0.98)	-
	Thyroid etiological diagnosis in adults.	Various conditions.	Pathological examination and follow up.	Clinical [d]	Physicians	100 (29/29)	-	BP	-	-	0.41 (0.24-0.61)	-
				Statistical	Oddie's model	100 (29/29)	-	BP	-	-	0.34 (0.19-0.54)	-
De Dombal (1972 & 1974)	Diagnosis in adults with acute abdominal pain.	Various conditions.	Final postoperative diagnosis or test results. Follow up.	Clinical [d]	Surgical house officers	100 (514/514)	-	BP	-	-	0.71 (0.67-0.75)	-
					Registrars and consultant surgeons	100 (552/552)	-	BP	-	-	0.80 (0.77-0.83)	-
				Statistical	Leeds' decision support system †	100 (552/552)	-	BP	-	-	0.91 (0.89-0.94)	-
Reale et al (1968)	Diagnosis of congenital cardiac diseases in children.	Various conditions.	Cardiac catheterization, surgery and autopsy.	Clinical [d]	1 cardiologist	100 (121/121)	-	BP	-	-	0.73 (0.64-0.80)	-
				Statistical	Reale's model †	100 (121/121)	-	BP	-	-	0.62 (0.53-0.71)	-

Abbreviations and notes: GI: Gastro Intestinal. AA: Acute Appendicitis. MI: Myocardial Infarction. ACS: Acute Coronary Syndrome. PE: Pulmonary Embolism. VTE: Venous Thromboembolic event. DVT: Deep Venous Thrombosis, ESC: European Society of Cardiology. ACC: American College of Cardiology. ICU: Intensive Care Unit. SICU: Surgical Intensive Care Unit. NICU: Neonatal Intensive Care Unit. PICU: Paediatric Intensive Care Unit. LAP: Lower Abdominal Pain. EM: Emergency Medicine. IM: Internal Medicine. ICM: Intensive Care Medicine. PM: Pulmonary Medicine. LR: Logistic Regression. ANN: Artificial Neural Networks. Cut offs: I. Cut-off implicit in prediction. U: Unclear. FD: First Diagnosis. PA: Positive answer. AVS (Any abnormal vital sign) Mod + (Moderate and severe). L + (Likely and yes) MH: Moderate-High).
[1] Independent sample. † Model derived for the same type of patients with sample from the same centre. [a] Prediction about the probability of the outcome. [b] Classificatory prediction for a dichotomous outcome. [c] Classificatory prediction for ordinal outcomes. [d] Classificatory prediction for set of nominal outcomes.

6.8. References

Adams, S.T. and Levenson, S.H. (2012). Clinical predictions rules. *BMJ.* 344:d8312.

Alberg, A.J., Park, J.W., Hager, B.W., Brock, M.V. and Diener-West, M. (2004). The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. *J Gen Intern Med*. 19(5 Pt 1):460-5.

Al Omar, M.Z. and Baldwin, G.A. (2002). Reappraisal of use of X-rays in childhood ankle and midfoot injuries. *Emerg Radiol*. 9(2):88-92.

Attia, M.W., Zaoutis, T., Klein, J.D., Meier, F.A. (2001). Performance of a predictive model for streptococcal pharyngitis in children. *Arch Pediatr Adolesc Med*. 155(6):687-91.

Baxt, W.G. (1991). Use of an artificial neural network for the diagnosis of myocardial infarction. *Ann Intern Med*. 115(11):843-8.

Bankowitz, R.A., McNeil, M.A., Challinor, S.M., Parker, R.C., Kapoor, W.N. and Miller, R.A. (1989). A computer-assisted medical diagnostic consultation service. Implementation and prospective evaluation of a prototype. *Ann Intern Med*. 110(10):824–32.

Beynon, R., Leeflang, M.M., McDonald, S., Eisinga, A., Mitchell, R.L., Whiting, P. and Glanville, J.M. (2013). Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. *Cochrane Database Syst Rev.* (9):MR000022.

Beyth, R.J., Quinn, L.M. and Landefeld, C.S. (1998). Prospective evaluation of an index for predicting the risk of major bleeding in outpatients treated with warfarin. *Am J Med.* 105(2):91-9.

Bigaroni, A., Perrier, A. and Bounameaux, H. (2000). Is clinical probability assessment of deep vein thrombosis by a score really standardized? *Thromb Haemost.* 83(5):788-9.

Blättler, W., Martinez, I. and Blättler, I.K. (2004). Diagnosis of deep venous thrombosis and alternative diseases in symptomatic outpatients. *Eur J Intern Med*. 15(5):305-11.

Bleeker, S.E., Moll, H.A., Steyerberg, E.W., Donders, A.R., Derksen-Lubsen, G., Grobbee, D.E. and Moons, K.G. (2003). External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol.* 56(9):826-32.

Bojang, K.A., Obaro, S., Morison, L.A. and Greenwood, B.M. (2000). A prospective evaluation of a clinical algorithm for the diagnosis of malaria in Gambian children. *Trop Med Int Health.* 5(4):231–6.

Borenstein, M., Hedges, L.V., Higgins, J.P. and Rothstein, H.R. (2009). *Critics of Meta-analysis.* In: *Introduction to meta-analysis.* Chichester, John Wiley & Sons Ltd. pp. 325-6.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D. and de Vet, H.C.; Standards for Reporting of Diagnostic Accuracy. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clin Chem*. 49(1):1-6.

Brannen, A.L. 2nd, Godfrey, L.J. and Goetter, W.E. (1989). Prediction of outcome from critical illness. A comparison of clinical judgment with a prediction rule. *Arch Intern Med*. 149(5):1083-6.

Brillman, J.C., Doezema, D., Tandberg, D., Sklar, D.P., Davis, K.D., Simms, S. and Skipper, B.J. (1996). Triage: limitations in predicting need for emergent care and hospital admission. *Ann Emerg Med*. 27(4):493-500.

Brokaw, J.P., Walker, W.C., Cifu, D.X. and Gardner, M. (2004). Sitting and standing tolerance in patients with chronic back pain: comparison between physician prediction and covert observation. *Arch Phys Med Rehabil.* 85(5):837-9.

Bruins Slot, M.H., Rutten, F.H., van der Heijden, G.J., Geersing, G.J., Glatz, J.F. and Hoes, A.W. (2011). Diagnosing acute coronary syndrome in primary care: comparison of the physicians' risk estimation and a clinical decision rule. *Fam Pract*. 28(3):323–8.

Carrier, M., Wells, P.S. and Rodger, M.A. (2006). Excluding pulmonary embolism at the bedside with low pre-test probability and D-dimer: safety and clinical utility of 4 methods to assign pre-test probability. *Thromb Res.* 117(4):469-74.

Chang, R.W., Lee, B., Jacobs, S. and Lee, B. (1989). Accuracy of decisions to withdraw therapy in critically ill patients: clinical judgment versus a computer model. *Crit Care Med.* 17(11):1091-7.

Chan, W.S., Lee, A., Spencer, F.A., Crowther, M., Rodger, M., Ramsay, T. and Ginsberg, J.S. (2009). Predicting deep venous thrombosis in pregnancy: out in "LEFt" field? *Ann Intern Med*. 151(2):85-92.

Chew, D.P., Junbo, G., Parsonage, W., Kerkar, P., Sulimov, V.A., Horsfall, M., Mattchoss, S.; Perceived Risk of Ischemic and Bleeding Events in Acute Coronary Syndrome Patients (PREDICT) Study Investigators. (2013). Perceived risk of ischemic and bleeding events in acute coronary syndromes. *Circ Cardiovasc Qual Outcomes*. 6(3):299-308.

Chew, D.P., Juergens, C., French, J., Parsonage, W., Horsfall, M., Brieger, D. and Quinn, S.; Predict study Investigators. (2014). An examination of clinical intuition in risk assessment among acute coronary syndromes patients: observations from a prospective multi-center international observational registry. *Int J Cardiol.* 171(2):209-16.

Christakis, N.A. and Iwashyna, T.J. (1998). Attitude and self-reported practice regarding prognostication in a national sample of internists. *Arch Intern Med.* 158(21):2389-95.

Collins, G.S., de Groot, J.A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L.M., Moons, K.G. and Altman, D.G. (2014). External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol.* 14:40. doi: 10.1186/1471-2288-14-40.

Conway, R., O'Riordan, D. and Silke, B. (2014). Consultant experience as a determinant of outcomes in emergency medical admissions. *Eur J Intern Med.* 25(2):151-5

Conway, R., Byrne, D.G., O'Riordan, D. and Silke, B. (2015). Improved outcomes of high-risk emergency medical admissions cared for by experienced physicians. *QJM.* 108(2):119-25.

Cook, N.R. (2008). Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem*. 54(1):17-23.

Cooper, J., Kapur, N. and Mackway-Jones, K. (2007). A comparison between clinicians' assessment and the Manchester Self-Harm Rule: a cohort study. *Emerg Med J*. 24(10):720-1.

Cornuz, J., Ghali, W.A., Hayoz, D., Stoianov, R., Depairon, M. and Yersin, B. (2002). Clinical prediction of deep venous thrombosis using two risk assessment methods in combination with rapid quantitative D-dimer testing. *Am J Med*. 112(3):198-203.

Crowe, L., Anderson, V. and Babl, F.E. (2010). Application of the CHALICE clinical prediction rule for intracranial injury in children outside the UK: impact on head CT rate. *Arch Dis Child*. 95:1017-22.

Dawes, R.M., Faust, D. and Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science.* 243(4899):1668-74.

Dawson, N.V. and Arkes, H.R. (1987). Systematic errors in medical decision making: judgment limitations. *J Gen Intern Med.* 2(3):183-7.

de Dombal, F.T., Leaper, D.J., Staniland, J.R., McCann, A.P. and Horrocks, J.C. (1972). Computer-aided diagnosis of acute abdominal pain. *Br Med J.* 2(5804):9-13.

De Dombal, F.T., Leaper, D.J., Horrocks, J.C., Staniland, J.R. and McCann, A.P. (1974). Human and computer-aided diagnosis of abdominal pain: Further report with emphasis on performance of clinicians. *Br Med J*. 1(5904):376-80.

de Dombal, F.T., Clamp, S.E., Leaper, D.J., Staniland, J.R. and Horrocks, J.C. (1975). Computer-aided diagnosis of lower gastrointestinal tract disorders. *Gastroenterology.* 68(2):252-60.

de Groot, N., van Oijen, M., Kessels, K., Hemmink, M., Weusten, B., Timmer, R., Hazen, W., van Lelyveld, N., Vermeijden, Jr., Curvers, W., Baak, L., Verburg, R., Bosman, J., de Wijkerslooth, L., de Rooij, J., Venneman, N., Pennings, M., van Hee, K., Scheffer, R., van Eijk, R., Meiland, R., Siersema, P. and Bredenoord, A. (2014). Prediction scores or gastroenterologists' gut feeling for triaging patients that present with acute upper gastrointestinal bleeding. *United European Gastroenterol J*. 2(3):197-205.

de Melo, S.W. Jr, Bhore, R. and Rockey, D.C. (2013). Clinical judgment does not circumvent the need for diagnostic endoscopy in upper gastrointestinal hemorrhage. *J Investig Med.* 61(8):1146-51.

Detrano R, Bobbio M, Olson H, Shandling, A., Ellestad, M.H., Alegria, E., Martinez-Caro, D., Righetti, A., Janosi, A. and Steinbrunn, W. (1992). Computer probability estimates of angiographic coronary artery disease: transportability and comparison with cardiologists' estimates. *Comput Biomed Res*. 25(5):468-85.

Dionne, C.E., Le Sage, N., Franche, R.L., Dorval, M., Bombardier, C. and Deyo, R.A. (2011). Five questions predicted long-term, severe, back-related functional limitations: evidence from three large prospective studies. *J Clin Epidemiol.* 64(1):54-66.

Dretzke, J., Ensor, J., Bayliss, S., Hodgkinson, J., Lordkipanidzé, M., Riley, R.D., Fitzmaurice, D. and Moore D. (2014). Methodological issues and recommendations for systematic reviews of prognostic studies: an example from cardiovascular disease. *Syst Rev.* 3:140.

Easter, J.S., Bakes, K., Dhaliwal, J., Miller, M., Caruso, E. and Haukoos, J.S. (2014). Comparison of PECARN, CATCH, and CHALICE rules for children with minor head injury: a prospective cohort study. *Ann Emerg Med.* 64(2):145-52.

Eddy, D.M. (1990). The Challenge. *JAMA.* 263(2):287-90.

Ægisdóttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook R.S., Nichols, C.N., Lampropoulos, G.K., Walker, B.S., Cohen, G. and Rush, J.D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist.* 34(3):341-82.

El-Solh, A.A., Hsiao, C.B., Goodnough, S., Serghani, J. and Grant, B.J. (1999). Predicting active pulmonary tuberculosis using an artificial neural network. *Chest.* 116(4):968-73.

Elstein, A.S. (1999). Heuristics and biases: selected errors in clinical reasoning. *Acad Med.* 74(7):791-4.

Emerman ,C.L., Dawson, N., Speroff, T., Siciliano, C., Effron, D., Rashad, F., Shaw, Z. and Bellon, E.L. (1991). Comparison of physician judgment and decision aids for ordering chest radiographs for pneumonia in outpatients. *Ann Emerg Med*. 20(11):1215-9.

Evenson, R.C., Altman, H., Sletten, I.W. and Cho, D.W. (1975).Accuracy of actuarial and clinical predictions for length of stay and unauthorized absence. *Dis Nerv Syst*. 36(5):250-2

Farion, K.J., Wilk, S., Michalowski, W., O'Sullivan, D. and Sayyad-Shirabad, J. (2013). Comparing predictions made by a prediction model, clinical score, and physicians: pediatric asthma exacerbations in the emergency department. *Appl Clin Inform*. 4(3):376-91.

Geersing, G.J., Janssen, K.J., Oudega, R., van Weert, H., Stoffers, H., Hoes, A. and Moons, K.; AMUSE Study Group. (2010). Diagnostic classification in patients with suspected deep venous thrombosis: physicians' judgement or a decision rule? *Br J Gen Pract*. 60(579):742-8.

Gerestein, C.G., van der Spek, D.W., Eijkemans, M.J., Bakker, J., Kooi, G.S. and Burger, C.W. (2009). Prediction of residual disease after primary cytoreductive surgery for advanced-stage ovarian cancer: accuracy of clinical judgment. *Int J Gynecol Cancer.* 19(9):1511-5

Gilovich, T., Griffin, D. And Kahneman, D. (2003). *Heuristics and Biases: The Psychology of Intuitive Judgment.* Cambridge, Cambridge University Press.

Glas, A.S., Pijnenburg, B.A., Lijmer, J.G., Bogaard, K., de, R.M., Keeman, J.N., Butzelaar, R.M. and Bossuyt, P.M. (2002). Comparison of diagnostic decision rules and structured data collection in assessment of acute ankle injury. *CMAJ*. 166(6):727-33.

Gottlieb, S.S. (2009). Prognostic indicators: useful for clinical care? *J Am Coll Cardiol*. 53(4):343-4.

Greenland, S. and O'Rourke, K. (2008). *Meta-analysis*. In *Modern Epidemiology*. K.J. Rothman, S. Greenland and T.L. Lash (eds.). Philadelphia, Lippincott Williams & Wilkins. pp.680-1.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E. and Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychol Assess.* 12(1):19-30.

Guyatt, G.H., Oxman, A.D., Schünemann, H.J., Tugwell, P. and Knottnerus, A. (2011). GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol.* 64(4):380-2.

Hadjianastassiou, V.G., Franco, L., Jerez, J.M., Evangelou, I.E., Goldhill, D.R., Tekkis, P.P. and Hands, L.J. (2006). Informed prognosis [corrected] after abdominal aortic aneurysm repair using predictive modeling techniques [corrected]. *J Vasc Surg.* 43(3):467-73.

Hallan, S., Asberg, A. and Edna, T.H. (1997). Estimating the probability of acute appendicitis using clinical criteria of a structured record sheet: the physician against the computer. *Eur J Surg*. 163(6):427-32.

Hausner, E., Guddat, C., Hermanns, T., Lampert, U. and Waffenschmidt, S. (2015). Development of search strategies for systematic reviews: validation showed the noninferiority of the objective approach. *J Clin Epidemiol.* 68(2):191-9.

Horrocks, J.C. and de Dombal, F.T. (1975). Computer-aided diagnosis of "dyspepsia". *Am J Dig Dis.*20:397-406.

Howitz, J. (2011). *The Philosophy of Evidence-based Medicine.* Oxford, Wiley-Blackwell.

Ikonen, J.K., Rokkanen, P.U., Grönroos, P., Kataja, J.M., Nykänen, P., de Dombal, F.T. and Softley, A. (1983). Presentation and diagnosis of acute abdominal pain in Finland: A computer aided study. *Ann Chir Gynaecol*. 72(6):332-6.

Jain, R., Duval, S. and Adabag, S. (2014). How accurate is the eyeball test?: a comparison of physician's subjective assessment versus statistical methods in estimating mortality risk after cardiac surgery. *Circ Cardiovasc Qual Outcomes*. 7(1):151–6.

Kabrhel, C., McAfee, A.T. and Goldhaber, S.Z. (2005). The contribution of the subjective component of the Canadian Pulmonary Embolism Score to the overall score in emergency department patients. *Acad Emerg Med*. 12(10):915-20.

Kabrhel, C., Mark Courtney, D., Camargo, C.A. Jr, Moore, C.L., Richman, P.B., Plewa, M.C., Nordenholtz, K.E., Smithline, H.A., Beam, D.M., Brown, M.D. and Kline, J.A. (2009). Potential impact of adjusting the threshold of the quantitative D-dimer based on pretest probability of acute pulmonary embolism. *Acad Emerg Med*. 16(4):325-32.

Kahneman, D. and Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *Am Psychol*. 64(6):515-26.

McClish, D.K. and Powell, S.H. (1989). How well can physicians estimate mortality in a medical intensive care unit? *Med Decis Making*. 9(2):125-32.

Keogh, C., Wallace, E., O'Brien, K.K., Galvin, R., Smith, S.M., Lewis, C., Cummins, A., Cousins, G., Dimitrov, B.D. and Fahey, T. (2014). Developing an international register of clinical prediction rules for use in primary care: a descriptive analysis. *Ann Fam Med.* 12(4):359-66.

Kirkeby, O.J and Riso, C. (1987). Use of a computer system for diagnosing acute abdominal pain in a small hospital. *Scand J Gastroenterol Suppl*. 128:174–6.

Kline, J.A., Courtney, D.M., Kabrhel, C., Moore, C.L., Smithline, H.A., Plewa, M.C., Richman, P.B., O'Neil, B.J. and Nordenholz, K. (2008). Prospective multicenter evaluation of the pulmonary embolism rule-out criteria. *J Thromb Haemost.* 6(5):772-80.

Kong, D.F., Lee, K.L., Harrell, F.E. Jr, Boswick, J.M., Mark, D.B., Hlatky, M.A., Califf, R.M. and Pryor, D.B. (1989). Clinical experience and predicting survival in coronary disease. *Arch Intern Med.* 149(5):1177-81.

Kong, V.Y., van der Linde, S., Aldous, C., Handley, J.J. and Clarke, D.L. (2014). The accuracy of the Alvarado score in predicting acute appendicitis in the black South African population needs to be validated. *Can J Surg.* 57(4):E121-5.

Kruse, J.A., Thill-Baharozian, M.C. and Carlson, R.W. (1988). Comparison of clinical assessment with APACHE II for predicting mortality risk in patients admitted to a medical intensive care unit. *JAMA.* 260(12):1739-42.

Laurent, M., Fournet, M., Feit, B., Oger, E., Donal, E., Thébault, C., Biron, Y., Beneux, X., Sellin, M., Le Reveillé, S., Flecher, E. and Leguerrier, A. (2013). Simple bedside clinical evaluation versus established scores in the estimation of operative risk in valve replacement for severe aortic stenosis. *Arch Cardiovasc Dis*. 106(12):651-60.

Laupacis, A., Sekar, N. and Stiell, I.G. (1997). Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA.* 277(6):488-94

Leibovici, L., Greenshtain, S., Cohen, O., Mor, F. and Wysenbeek, A.J. (1991). Bacteremia in febrile patients. A clinical model for diagnosis. *Arch Intern Med.* 151(9):1801-6.

Leung, K.M., Hopman, W.M. and Kawakami, J. (2012). Challenging the 10-year rule: The accuracy of patient life expectancy predictions by physicians in relation to prostate cancer management. *Can Urol Assoc J.* 6(5):367-73.

Lintula, H., Kokki, H., Kettunen, R. and Eskelinen, M. (2009). Appendicitis score for children with suspected appendicitis. A randomized clinical trial. *Langenbecks Arch Surg*. 394(6):999-1004.

Lintula, H., Kokki, H., Pulkkinen, J., Kettunen, R., Gröhn, O. and Eskelinen, M. (2010). Diagnostic score in acute appendicitis. Validation of a diagnostic score (Lintula score) for adults with suspected appendicitis. *Langenbecks Arch Surg*. 395(5):495-500.

Litton, E., Ho, K.M. and Webb, S.A. (2012). Comparison of physician prediction with 2 prognostic scoring systems in predicting 2-year mortality after intensive care admission: a linked-data cohort study. *J Crit Care*. 27(4):423.e9-15.

Mán, E., Simonka, Z., Varga, A., Rárosi, F. and Lázár, G. (2014). Impact of the Alvarado score on the diagnosis of acute appendicitis: comparing clinical judgment, Alvarado score, and a new modified score in suspected appendicitis: a prospective, randomized clinical trial. *Surg Endosc*. 28(8):2398-405.

Marcin, J.P., Pollack, M.M., Patel, K.M., Sprague, B.M. and Ruttimann, U.E. (1999). Prognostication and certainty in the pediatric intensive care unit. *Pediatrics.* 104(4 Pt 1):868-73.

Marcin, J.P., Pollack, M.M., Patel, K.M. and Ruttimann, U.E. (2000). Combining physician's subjective and physiology-based objective mortality risk predictions. *Crit Care Med.* 28(8):2984-90.

Maguire, J.L., Kulik, D.M., Laupacis, A., Kuppermann, N., Uleryk, E.M. and Parkin, P.C. (2011). Clinical prediction rules for children: a systematic review. *Pediatrics.* 128(3):e666-77.

Mallett, S., Royston, P., Dutton, S., Waters, R. and Altman, D.G. (2010). Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med*. 8:20

Mallett, S., Halligan, S., Thompson, M., Collins, G.S. and Altman, D.G. (2012). Interpreting diagnostic accuracy studies for patient care. *BMJ*. 345:e3999.

Marchese, M.C. (1992). Clinical versus actuarial prediction: a review of the literature. *Percept Mot Skills.* 75(2):583-94.

McGinn, T., Jervis, R., Wisnivesky, J., Keitz, S. and Wyer, P.C.; Evidence-based Medicine Teaching Tips Working Group. (2008). Tips for teachers of evidence-based medicine: clinical prediction rules (CPRs) and estimating pretest probability. *J Gen Intern Med.* 23(8):1261-8.

Meehl, P.E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, University of Minnesota.

Meltzer, A.C., Baumann, B.M., Chen, E.H., Shofer, F.S. and Mills, A.M. (2013). Poor sensitivity of a modified Alvarado score in adults with suspected appendicitis. *Ann Emerg Med*. 62(2):126–31.

Meyer, A.A., Messick, W.J., Young, P., Baker, C.C., Fakhry, S., Muakkassa, F., Rutherford, E.J., Napolitano, L.M. and Rutledge, R. (1992). Prospective comparison of clinical judgment and APACHE II score in predicting the outcome in critically ill surgical patients. *J Trauma*. 32(6):747-53; discussion 753-4.

Mitchell, A.M., Garvey, J.L., Chandra, A., Diercks, D., Pollack, C.V. and Kline, J.A. (2006). Prospective multicenter study of quantitative pretest probability assessment to exclude acute coronary syndrome for patients evaluated in emergency department chest pain units. *Ann Emerg Med*. 47(5):447.

Moons, K.G., Donders, A.R., Steyerberg, E.W. and Harrell, F.E. (2004). Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol.* 57(12):1262-70.

Moons, K.G., Altman, D.G., Reitsma, J.B., Ioannidis, J.P., Macaskill, P., Steyerberg, E.W., Vickers, A.J., Ransohoff, D.F. and Collins, G.S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 162(1):W1-73.

Oddie, T.H., Hales, I.B., Stiel, J.N., Reeve, T.S., Hooper, M., Boyd, C.M. and Fisher, D.A. (1974). Prospective trial of computer program for diagnosis of thyroid disease. *J Clin Endocrinol Metab*. 38(5):876-82.

Penaloza, A., Verschuren, F., Dambrine, S., Zech, F., Thys, F. and Roy, P.M. (2012). Performance of the Pulmonary Embolism Rule-out Criteria (the PERC rule) combined with low clinical probability in high prevalence population. *Thromb Res*. 129(5):e189–93.

Penaloza, A., Verschuren, F., Meyer, G., Quentin-Georget, S., Soulie, C., Thys, F. and Roy, P.M. (2013). Comparison of the unstructured clinician gestalt, the wells score, and the revised Geneva score to estimate pretest probability for suspected pulmonary embolism. *Ann Emerg Med*. 62(2):117–124.e2.

Perel, P., Edwards, P., Wentz, R. and Roberts, I. (2006). Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak.* 6:38.

Phillips, C.B., Sackett, D., Badenoch, D., Straus, S., Haynes, B. and Dawes, M. (1998). *"CEBM Levels of Evidence"* Retrieved 01 January, 2016, from http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/

Pons, J.M., Borras, J.M., Espinas, J.A., Moreno, V., Cardona, M. and Granados, A. (1999). Subjective versus statistical model assessment of mortality risk in open heart surgical procedures. *Ann Thorac Surg.* 67(3):635-40.

Poretsky, L., Leibowitz, I.H. and Friedman, S.A. (1985). The diagnosis of myocardial infarction by computer-derived protocol in a municipal hospital. *Angiology*. 36(3):165-70.

Pruekprasert, P., Maipang, T., Geater, A., Apakupakul, N. and Ksuntigij, P. (2004). Accuracy in diagnosis of acute appendicitis by comparing serum C-reactive protein measurements, Alvarado score and clinical impression of surgeons. *J Med Assoc Thai.* 87(3):296-303.

Quinn, J.V., Stiell, I.G., McDermott, D.A., Kohn, M.A. and Wells, G.A. (2005). The San Francisco Syncope Rule vs physician judgment and decision making. *Am J Emerg Med*. 23(6):782-6.

Rabin, A., Shashua, A., Pizem, K., Dickstein, R. and Dar, G. (2014). A clinical prediction rule to identify patients with low back pain who are likely to experience short-term success following lumbar stabilization exercises: a randomized controlled validation study. *J Orthop Sports Phys Ther.* 44(1):6-B13.

Reale A, Maccacaro GA, Rocca E, D'Intino, S., Gioffre, P.A., Vestri, A. and Motolese, M. (1968). Computer diagnosis of congenital heart disease. *Comput Biomed Res.* 1(6):533-49.

Richardson, S., Khan, S., McCullagh, L., Kline, M., Mann, D. and McGinn, T. (2015). Healthcare provider perceptions of clinical prediction rules. *BMJ Open.* 5(9):e008461.

Richman, P.B., McCuskey, C.F., Nashed, A., Fuchs, S., Petrik, R., Imperato, M. and Hollander, J.E. (1997). Performance of two clinical decision rules for knee radiography. *J Emerg Med.* 15(4):459-63.

Sampson, M., McGowan, J., Cogo, E., Grimshaw, J., Moher, D. and Lefebvre, C. (2009). An evidence-based practice guideline for the peer review of electronic search strategies. *J Clin Epidemiol.* 62(9):944-52.

Sanson, B.J., Lijmer, J.G., Mac Gillavry, M.R., Turkstra, F., Prins, M.H. and Büller, H.R. (2000). Comparison of a clinical probability estimate and two clinical models in patients with suspected pulmonary embolism. ANTELOPE-Study Group. *Thromb Haemost.* 83(2):199-203.

Scholz, N., Bäsler, K., Saur, P., Burchardi, H. and Felder, S. (2004). Outcome prediction in critical care: Physicians' prognoses vs scoring systems. *Eur J Anaesthesiol*. 21(8):606-11.

Shojania, K.G., Burton, E.C., McDonald, K.M., and Goldman, L. (2003). Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. *JAMA.* 289(21):2849-56

Singh, S., Nosyk, B., Sun, H., Christenson, J.M., Innes, G. and Anis, A.H. (2008). Value of information of a clinical prediction rule: informing the efficient use of healthcare and health research resources. *Int J Technol Assess Health Care.* 24(1):112-9.

Siontis, G.C., Tzoulaki, I., Siontis, K.C. and Ioannidis, J.P. (2012). Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ.* 344:e3318.

Siontis, G.C., Tzoulaki, I., Castaldi, P.J. and Ioannidis, J.P. (2015). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol.* 68(1):25-34.

Smith, L., Gilhooly, K. and Walker, A. (2003). Factors influencing prescribing decisions in the treatment of depression: a social judgement theory approach. *Applied Cognitive Psychology.* 17(1):51-63.

Sniderman, A.D., D'Agostino, R.B. Sr and Pencina, M.J. (2015). The Role of Physicians in the Era of Predictive Analytics. *JAMA.* 314(1):25-6.

Spengler, P.M. and Pilipis, L.A. (2015). A comprehensive meta-reanalysis of the robustness of the experience-accuracy effect in clinical judgment. *J Couns Psychol*. 62(3):360-78.

Stein, J., Louie, J., Flanders, S., Maselli, J., Hacker, J.K., Drew, W.L. and Gonzales, R. (2005). Performance characteristics of clinical diagnosis, a clinical decision rule, and a rapid influenza test in the detection of influenza infection in a community sample of adults. *Ann Emerg Med*. 46(5):412–9.

Stevens, S.M., Richardson, D.K., Gray, J.E., Goldmann, D.A. and McCormick, M.C. (1994). Estimating neonatal mortality risk: an analysis of clinicians' judgments. *Pediatrics.* 93(6 Pt 1):945-50.

Steyerberg, E.W. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* New York, Springer.

Steyerberg, E.W., Moons, K.G., van der Windt, D.A., Hayden, J.A., Perel, P., Schroter, S., Riley, R.D., Hemingway, H. and Altman, D.G. PROGRESS Group. (2013). Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med.* 10(2):e1001381.

Sutton, G.C. (1989). How accurate is computer-aided diagnosis? *Lancet*. 2(8668):905–8.

Takwoingi, Y., Guo, B., Riley, R.D. and Deeks, J.J. (2015). Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. Stat Methods Med Res. pii: 0962280215592269.

Tenório, J.M., Hummel, A.D., Cohrs, F.M., Sdepanian, V.L., Pisa, I.T. and de Fátima Marin, H. (2011). Artificial intelligence techniques applied to the development of a decision-support system for diagnosing celiac disease. *Int J Med Inform*. 80(11):793-802.

Thompson, D.D., Murray, G.D., Sudlow, C.L., Dennis, M. and Whiteley, W.N. (2014). Comparison of statistical and clinical predictions of functional outcome after ischemic stroke. *PLoS One*. 9(10):e110189.

Van Calster, B., Vergouwe, Y., Looman, C.W., Van Belle, V., Timmerman, D. and Steyerberg, E.W. (2012). Assessing the discriminative ability of risk models for more than two outcome categories. *Eur J Epidemiol.* 27(10):761-70.

Van Calster, B., Steyerberg, E.W. and Harrell, F.H. (2015). Risk Prediction for Individuals. *JAMA.* 314(17):1875.

van Gerven, M.A., Jurgelenaite, R., Taal, B.G., Heskes, T. and Lucas, P.J. (2007). Predicting carcinoid heart disease with the noisy-threshold classifier. *Artif Intell Med*. 40(1):45-55.

Wang, B., Lin, Y., Pan, F.S., Yao, C., Zheng, Z.Y., Cai, D. and Xu. X.D. (2013). Comparison of empirical estimate of clinical pretest probability with the Wells score for diagnosis of deep vein thrombosis. *Blood Coagul Fibrinolysis*. 24(1):76-81.

Whiting, P.F., Rutjes, A.W., Westwood, M.E., Mallett, S., Deeks, J.J., Reitsma, J.B., Leeflang, M.M., Sterne, J.A. and Bossuyt, P.M.; QUADAS-2 Group. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 155(8):529-36.

Yap, C.H., Reid, C., Yii, M., Rowland, M.A., Mohajeri, M., Skillington, P.D., Seevanayagam, S. and Smith, J.A. Validation of the EuroSCORE model in Australia. *Eur J Cardiothorac Surg*. 29(4):441-6.

Main Conclusions

The central aim of this doctoral dissertation was to examine the role of clinical judgment in the care of individuals in the era of EBM. The EBM approach has de facto encouraged physicians to make clinical recommendations only when in line with the best research evidence, and has effectively decreased the room for clinical discretion.

Against this emphasis on EBM, I have argued that medicine ought to recognise that physicians' judgment has a pivotal role in clinical decision-making. The exercise of clinical judgment allows physicians to address what I have called the Problem of Extra Information by estimating the right probabilities for each patient.

The PEI arises because physicians are subtly and tacitly coerced not to pay proper attention to part of the information they have about the individual patients. This because they are trained to restrict themselves to recommendations supported by probabilities obtained by means of the best research evidence available.

I argued that there is a sense in which the probabilities favoured by EBM are the wrong probabilities for clinical recommendations: insofar as such probabilities ignore relevant information, they are not the probabilities with respect to which the physician ought to maximize expected utility for her patient. Given this problem, I proposed the Discretionary Approach: this was based on the ideas that (a) the objective probabilities of interest for clinical recommendations are the probabilities in the reference class defined by everything the physician knows about the patient, and (b) physicians standardly need to exercise their judgment to estimate these probabilities.

As I explained, research improvements have important merits, some of which ameliorate the PEI by narrowing the gap between research probabilities and the right probabilities for individual patients. Nonetheless, it would be erroneous to conclude that movements such as personalised medicine or precision medicine, which present themselves as providing truly individualised predictions and prescriptions, normally provide physicians with the right probabilities for clinical recommendations. Even after these movements have done their work, the PEI typically remains and requires the rational physician to exercise her best judgement to address it.

The empirical research presented in my final two chapters suggests that the DA would improve the care of individuals. Of course, the empirical evidence presented is not as direct as one would like, and has various limitations inherited from primary studies. Nonetheless, I

believe that the arguments and empirical data presented in this thesis is sufficiently robust to cast doubt on the assumed "optimality" of the EBM approach.

EBM has brought about positive changes in healthcare; however, the merits of EBM do not turn this approach into the "best choice available". Further changes might improve the care of patients even further, in particular if physicians learn better ways to integrate research with non-research sources of evidence.

It is worth emphasising that my aim in this thesis has not been to try to defend the role of clinical judgment in a blind way. I agree that there is some evidence that clinical judgment can be unreliable, and I certainly concur with those who claim that it is not a suitable method for generating general population-level medical knowledge. So, by calling for more room for the exercise of clinical judgment, I am not advocating a blind trust in clinical judgment, but rather urging a systematic study of its performance. One of the negative by-products of the EBM movement has been a lack of interest in learning about and improving clinical judgment.

What I am certainly arguing against, though, are any a priori assumptions that the EBM approach is always better than physicians' predictions and prescriptions. As I have shown, this faith is not supported by evidence.

Once we recognize the general presence of the PEI, and the consequent need for clinical judgment, it will become important to learn more about different levels of physicians' performance in a variety of real settings. Of course, any such analysis will face many practical challenges. But it is certainly better than assuming that clinical judgment is unreliable altogether and it should not be given any official place in the context of clinical inference. We need to be serious about understanding clinical practice, by measuring its success and failures in practice, and helping physicians to improve it as much as possible.

As the Kenneth Goodman (2003) suggests *"We should neither wallow in [clinical judgment] nor hope to overcome it. Rather, reducing conceptual biases, disclosure, and increased and improved learning are what we should consider as the core faculties that are enlisted when we [exercise our judgment]. This has a wonderful advantage over other vague formulations or hopeful stipulations: It provides grounds for explaining the good outcomes on those occasions when we get it right…and also, and more importantly, a framework for improvement."* (p.132).

The importance of exercising sound clinical judgment, the practical value of clinical experience, and the great significance of cultivating sensitivity to patients' values and preferences have all been officially embraced at some point by prominent supporters of EBM

(Sackett, et al. 1996). Despite this, EBM as it is practised seems to have forgotten the ideals envisaged by its founders. If EBM wants to return to its roots, then it should examine the practical import of considered experience. Clinical judgement, just as much as research evidence and standardization, needs to be studied and improved by medical bodies, governments and funding agencies.